Global Convergence of Stochastic Gradient Descent for Some Non-convex Matrix Problems

Christopher De Sa, Kunle Olukotun, and Christopher Ré
cdesa@stanford.edu, kunle@stanford.edu, chrismre@stanford.edu
Departments of Electrical Engineering and Computer Science
Stanford University, Stanford, CA 94309

February 11, 2015

Abstract

Stochastic gradient descent (SGD) on a low-rank factorization [9] is commonly employed to speed up matrix problems including matrix completion, subspace tracking, and SDP relaxation. In this paper, we exhibit a step size scheme for SGD on a low-rank least-squares problem, and we prove that, under broad sampling conditions, our method converges globally from a random starting point within $O(\epsilon^{-1} n \log n)$ steps with constant probability for constant-rank problems. Our modification of SGD relates it to stochastic power iteration. We also show experiments to illustrate the runtime and convergence of the algorithm.

1 Introduction

We analyze an algorithm to solve the stochastic optimization problem

$$\begin{array}{ll} \text{minimize} & \mathbf{E}\left[\left\|\tilde{A} - X\right\|_F^2\right] \\ \text{subject to} & X \in \mathbb{R}^{n\times n}, \mathbf{rank}\left(X\right) \leq p, X \succeq 0, \end{array} \tag{1}$$

where p is an integer and $\tilde{A}$ is a symmetric matrix drawn from some distribution with bounded covariance. The solution to this problem is the matrix formed by zeroing out all but the largest p eigenvalues of the matrix $\mathbf{E}[\tilde{A}]$. This problem, or problems that can be transformed to this problem, appears in a variety of machine learning applications including matrix completion [14, 25, 36], general data analysis [37], subspace tracking [6], principle component analysis [3], optimization [10, 23, 27, 29], and recommendation systems [20, 32].

Sometimes, (1) arises under conditions in which the samples $\tilde{A}$ are sparse, but the matrix X would be too large to store and operate on efficiently; a standard heuristic to use in this case is a low-rank factorization [9]. The idea is to substitute $X = YY^T$ and solve the problem

$$\begin{array}{ll} \text{minimize} & \mathbf{E}\left[\left\|\tilde{A} - YY^T\right\|_F^2\right] \\ \text{subject to} & Y \in \mathbb{R}^{n\times p}. \end{array} \tag{2}$$

By construction, if we set $X = YY^T$, then $X \in \mathbb{R}^{n\times n}$, $\mathbf{rank}\left(X\right) \leq p$, and $X \succeq 0$; this allows us to drop these constraints. Instead of having to store the matrix X (of size n^2), we only need to store the matrix Y (of size np).

In practice, many people use stochastic gradient descent (SGD) to solve (2). Efficient SGD implementations can scale to very large datasets [2, 7, 8, 16, 24, 30, 33, 36]. However, standard stochastic gradient descent on (2) does not converge globally, in the sense that there will always be some initial values for which the norm of the iterate will diverge (see Appendix A).

People have attempted to compensate for this with sophisticated methods like geodesic step rules [27] and manifold projections [1]; however, even these methods cannot guarantee global convergence. Motivated by this, we describe Alecton, an algorithm for solving (2), and analyze its convergence. Alecton is an SGD-like algorithm that has a simple update rule with a step size that is a simple function of the norm of the iterate Y_k. We show that Alecton converges globally. We make the following contributions:

- We establish the convergence rate to a global optimum of Alecton using a random initialization; in contrast, prior analyses [11, 25] have required more expensive initialization methods, such as the singular value decomposition of an empirical average of the data.
- In contrast to previous work that uses bounds on the magnitude of the noise [21], our analysis depends only on the variance of the samples. As a result, we are able to be robust to different noise models, and we apply our technique to these problems, which did not previously have global convergence rates:
  - *matrix completion*, in which we observe entries of A one at a time [25, 28] (Section 4.1),
  - *phase retrieval*, in which we observe $\mathrm{tr}(u^T A v)$ for randomly selected u, v [11, 13] (Section 4.3), and
  - *subspace tracking*, in which A is a projection matrix and we observe random entries of a random vector in its column space [6] (Section 4.4).

  Our result is also robust to different noise models.
- We describe a martingale-based analysis technique that is novel in the space of non-convex optimization. We are able to generalize this technique to some simple regularized problems, and we are optimistic that it has more applications.

1.1 Related Work

Much related work exists in the space of solving low-rank factorized optimization problems. Foundational work in this space was done by Burer and Monteiro [9, 10], who analyzed the low-rank factorization of general semidefinite programs. Their results focus on the classification of the local minima of such problems, and on conditions under which no non-global minima exist. They do not analyze the convergence rate of SGD.

Another general analysis in Journée et al. [27] exhibits a second-order algorithm that converges to a local solution. Their results use manifold optimization techniques to optimize over the manifold of low-rank matrices. These approaches have attempted to correct for falling off the manifold using Riemannian retractions [27], geodesic steps [6], or projections back onto the manifold. General non-convex manifold optimization techniques [1] tell us that first-order methods, such as SGD, will converge to a fixed point, but they provide no convergence rate to the global optimum. Our algorithm only involves a simple rescaling, and we are able to provide global convergence results.

Our work follows others who have studied individual problems that we consider. Jain et al. [25] study matrix completion and provides a convergence rate for an exact recovery algorithm, alternating minimization. Candès et al. [11] provide a similar result for phase retrieval. In contrast to these results, which require expensive SVD-like operations to initialize, our results allow random initialization. Our provided convergence rates apply to additional problems and SGD algorithms that are used in practice (but are not covered

by previous analysis). However, our convergence rates are slower in their respective settings. This is likely unavoidable in our setting, as we show that our convergence rate is optimal in this more general setting (see Appendix E).

A related class of algorithms that are similar to Alecton is stochastic power iteration [3]. These algorithms reconsider (1) as an eigenvalue problem, and uses the familiar power iteration algorithm, adapted to a stochastic setting. Stochastic power iteration has been applied to a wide variety of problems [3, 26]. Oja [31] show convergence of this algorithm, but provides no rate. Arora et al. [4] analyze this problem, and state that "obtaining a theoretical understanding of the stochastic power method, or of how the step size should be set, has proved elusive." Our paper addresses this by providing a method for selecting the step size, although our analysis shows convergence for any sufficiently small step size.

Shamir [35] provide exponential-rate local convergence results for a stochastic power iteration algorithm for PCA. As they note, it can be used in practice to improve the accuracy of an estimate returned by another, globally-convergent algorithm such as Alecton.

Also recently, Balsubramani et al. [5] and Hardt and Price [21] provide a global convergence rate for the stochastic power iteration algorithm. Our result only depends on the variance of the samples, while both their results require absolute bounds on the magnitude of the noise. This allows us to analyze a different class of noise models, which enables us to do matrix completion, phase retrieval, and subspace tracking in the same model.

2 Algorithmic Derivation

We focus on the low-rank factorized stochastic optimization problem (2). We can rewrite the objective as $\mathbf{E}\left[\tilde{f}(Y)\right]$, with sampled objective function

$$\tilde{f}(Y) = \mathbf{tr}\left(YY^TYY^T\right) - 2\mathbf{tr}\left(Y\tilde{A}Y^T\right) + \left\|\tilde{A}\right\|_F^2.$$

In the analysis that follows, we let $A = \mathbf{E}\left[\tilde{A}\right]$, and let its eigenvalues be $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ with corresponding orthonormal eigenvectors $u_1, u_2, \ldots, u_n$ (such a decomposition is guaranteed since A is symmetric). The standard stochastic gradient descent update rule for this problem is, for some step size α_k,

$$\begin{aligned} Y_{k+1} &= Y_k - \alpha_k \nabla \tilde{f}_k(Y) \\ &= Y_k - 4\alpha_k \left(Y_k Y_k^T Y_k - \tilde{A}_k Y_k\right), \end{aligned}$$

where $\tilde{A}_k$ is the sample we use at timestep k.

The low-rank factorization introduces symmetry into the problem. If we let

$$\mathcal{O}_p = \left\{U \in \mathbb{R}^{p\times p} \mid U^TU = I_p\right\}$$

denote the set of orthogonal matrices in $\mathbb{R}^{p\times p}$, then $\tilde{f}(Y) = \tilde{f}(YU)$ for any $U \in \mathcal{O}_p$. Previous work has used manifold optimization techniques to solve such symmetric problems [27]. Absil et al. [1] state that stochastic gradient descent on a manifold has the general form

$$x_{k+1} = x_k - \alpha_k G_{x_k}^{-1} \nabla \tilde{f}_k(x_k),$$

where G_x is the matrix such that for all u and v,

$$u^T G_x v = \langle u, v\rangle_x,$$

where the right side of this equation denotes the *Riemannian metric* [15] of the manifold at x. For (2), the manifold in question is

$$\mathcal{M} = \mathbb{R}^{n\times p}/\mathcal{O}_p,$$

which is the quotient manifold of $\mathbb{R}^{n\times p}$ under the orthogonal group action. According to Absil et al. [1], this manifold has induced Riemannian metric

$$\langle U, V\rangle_Y = \mathbf{tr}\left(UY^TYV^T\right). \tag{3}$$

For Alecton, we are free to pick any Riemannian metric and step size. Inspired by (3), we pick a new step size parameter η, and let $\alpha_k = \frac{1}{4}\eta$ and set

$$\langle U, V\rangle_Y = \mathbf{tr}\left(U(I+\eta Y^TY)V^T\right).$$

With this, the SGD update rule becomes

$$\begin{aligned} Y_{k+1} &= Y_k - \eta\left(Y_kY_k^TY_k - \tilde{A}_kY_k\right)\left(I+\eta Y_k^TY_k\right)^{-1} \\ &= \left(Y_k\left(I+\eta Y_k^TY_k\right) - \eta\left(Y_kY_k^TY_k - \tilde{A}_kY_k\right)\right)\left(I+\eta Y_k^TY_k\right)^{-1} \\ &= \left(I+\eta\tilde{A}_k\right)Y_k\left(I+\eta Y_k^TY_k\right)^{-1}. \end{aligned}$$

For $p = 1$, choosing a Riemannian metric to use with SGD results in the same algorithm as choosing an SGD step size that depends on the iterate Y_k. The same update rule would result if we substituted

$$\alpha_k = \frac{1}{4}\eta\left(1+\eta Y^TY\right)^{-1}$$

into the standard SGD update formula. We can think of this as the manifold results giving us intuition on how to set our step size.

The reason why selecting this particular step size/metric is useful in practice is that we can run the simpler update rule

$$\bar{Y}_{k+1} = \left(I+\eta\tilde{A}_k\right)\bar{Y}_k. \tag{4}$$

If $\bar{Y}_0 = Y_0$, the iteration will satisfy the property that the column space of Y_k will always be equal to the column space of $\bar{Y}_k$, (since $C(XY) = C(X)$ for any invertible matrix Y). That is, if we just care about computing the column space of Y_k, we can do it using the much simpler update rule (4). Intuitively, we have transformed an optimization problem operating in the whole space $\mathbb{R}^n$ to one operating on the Grassmannian; one benefit of Alecton is that we don't have to work on the actual Grassmannian, but get some of the same benefits from a rescaling of the Y_k space. In this specific case, the Alecton update rule is akin to stochastic power iteration, since it involves a repeated multiplication by the sample; this would not hold for optimization on other manifolds.

We can use (4) to compute the column space (or "angular component") of the solution, before then recovering the rest of the solution (the "radial component") using averaging. Doing this corresponds to Algorithm 1, Alecton. Notice that, unlike most iterative algorithms for matrix recovery, Alecton does not require any special initialization phase and can be initialized randomly.

Analysis Analyzing this algorithm is challenging, as the low-rank decomposition also introduces symmetrical families of fixed points. Not all these points are globally optimal: in fact, a fixed point will occur whenever

$$YY^T = \sum_{i\in C}\lambda_iu_iu_i^T$$

Algorithm 1 Alecton: Solve stochastic matrix problem

Require: $\eta \in \mathbb{R}$, $K \in \mathbb{N}$, $L \in \mathbb{N}$, and a sampling distribution $\mathcal{A}$

▷ ***Angular component (eigenvector) estimation phase***

Select Y_0 uniformly in $\mathbb{R}^{n \times m}$ s.t. $Y_0^T Y_0 = I$.

for $k = 0$ **to** $K - 1$ **do**

Select $\tilde{A}_k$ uniformly and independently at random from the sampling distribution $\mathcal{A}$.

$Y_{k+1} \leftarrow Y_k + \eta \tilde{A}_k Y_k$

end for

$\hat{Y} \leftarrow Y_K \left(Y_K^T Y_K\right)^{-\frac{1}{2}}$

▷ ***Radial component (eigenvalue) estimation phase***

$R_0 \leftarrow 0$

for $l = 0$ **to** $L - 1$ **do**

Select $\tilde{A}_l$ uniformly and independently at random from the sampling distribution $\mathcal{A}$.

$R_{l+1} \leftarrow R_l + \hat{Y}^T \tilde{A}_l \hat{Y}$

end for

$\bar{R} \leftarrow R_L / L$

return $\hat{y} \bar{R}^{\frac{1}{2}}$

for any set C of size less than p.

One consequence of the non-optimal fixed points is that the standard proof of SGD's convergence, in which we choose a Lyapunov function and show that this function's expectation decreases with time, cannot work. This is because, if such a Lyapunov function were to exist, it would show that no matter where we initialize the iteration, convergence to a global optimum will still occur rapidly; this cannot be possible due to the presence of the non-optimal fixed points. Thus, a standard statement of global convergence, that convergence occurs uniformly regardless of initial condition, cannot hold.

We therefore use martingale-based methods to show convergence. Specifically, our attack involves defining a process x_k with respect to the natural filtration $\mathcal{F}_k$ of the iteration, such that x_k is a supermartingale, that is $\mathbf{E}\left[x_{k+1} | \mathcal{F}_k\right] \leq x_k$. We then use the *optional stopping theorem* [17] to bound both the probability and rate of convergence of x_k, from which we derive convergence of the original algorithm. We describe this analysis in the next section.

3 Convergence Analysis

First, we need a way to define convergence for the angular phase. For most problems, we want $C(Y_k)$ to be as close as possible to the span of $u_1, u_2, \ldots, u_p$. However, for some cases, this is not what we want. For example, consider the case where $p = 1$ but $\lambda_1 = \lambda_2$. In this case, the algorithm could not recover u_1, since it is indistinguishable from u_2. Instead, it is reasonable to expect $C(Y_k)$ to converge to the span of u_1 and u_2.

To handle this case, we instead want to measure convergence to the subspace spanned by some number, $q \geq p$, of the algebraically largest eigenvectors (in most cases, $q = p$). For a particular q, let U be the projection matrix onto the subspace spanned by $u_1, u_2, \ldots, u_q$, and define Δ, the *eigengap*, as $\Delta = \lambda_q - \lambda_{q+1}$. We now let $\epsilon > 0$ be an arbitrary error term, and define an angular success condition for Alecton.

Definition 1. When running the angular phase of Alecton, we say that *success has occurred* at timestep k if and only if for all $z \in \mathbb{R}^p$,

$$\frac{\left\|U Y_k z\right\|^2}{\left\|Y_k z\right\|^2} \geq 1 - \epsilon.$$

This condition requires that all members of the column space of Y_k are close to the desired subspace. We say that *success has occurred by time* t if success has occurred for some timestep $k < t$. Otherwise, we say the algorithm has *failed*, and we let F_t denote this failure event.

To prove convergence, we need to put some restrictions on the problem. Our theorem requires the following three conditions.

Condition 1 (Alecton Variance)**.** *A sampling distribution $\mathcal{A}$ with expected value A satisfies the* Alecton Variance Condition *(AVC) with parameters (σ_a, σ_r) if and only if for any $y \in \mathbb{R}$ and for any symmetric matrix $W \succeq 0$ that commutes with A, if $\tilde{A}$ is sampled from $\mathcal{A}$, the following bounds hold:*

$$\mathbf{E}\left[y^T \tilde{A}^T W \tilde{A} y\right] \leq \sigma_a^2 \mathbf{tr}\left(W\right) \|y\|^2$$

and

$$\mathbf{E}\left[\left(y^T \tilde{A} y\right)^2\right] \leq \sigma_r^2 \|y\|^4 .$$

In Section 4, we show several models that satisfy AVC.

Condition 2 (Alecton Rank)**.** *An instance of Alecton satisfies the* Alecton Rank Condition *if either $p = 1$ (rank-1 recovery), or each sample $\tilde{A}$ from $\mathcal{A}$ is rank-1 (rank-1 sampling).*

Most of the noise models we analyze have rank-1 samples, and so satisfy the rank condition.

Condition 3 (Alecton Step Size)**.** *Define γ as*

$$\gamma = \frac{2n\sigma_a^2 p^2 (p+\epsilon)}{\Delta\epsilon}\eta.$$

This represents a constant step size parameter that is independent of problem scaling. An instance of Alecton satisfies the Alecton Step Size Condition *if and only if $\gamma \leq 1$.*

Note that the step size condition is only an upper bound on the step size. This means that, even if we do not know the problem parameters exactly, we can still choose a feasible step size as long as we can bound them. (However, smaller step sizes imply slower convergence, so it is a good idea to choose η as large as possible.)

We will now define a useful function, then state our main theorem that bounds the probability of failure.

Definition 2. For some p, let $R \in \mathbb{R}^{p \times p}$ be a random matrix the entries of which are independent standard normal random variables. Define function Z_p as

$$Z_p(\gamma) = 2\left(1 - \mathbf{E}\left[\left|I + \gamma p^{-1}(R^T R)^{-1}\right|^{-1}\right]\right).$$

Theorem 1. *Assume that we run an instance of Alecton that satisfies the variance, rank, and step size conditions. Then for any t, the probability that the angular phase will have failed up to time t is*

$$P\left(F_t\right) \leq Z_p(\gamma) + \frac{4n\sigma_a^2 p^2 (p+\epsilon)}{\Delta^2 \gamma \epsilon t} \log\left(\frac{np^2}{\gamma q \epsilon}\right). \tag{5}$$

Also, in the radial phase, for any constant ψ it holds that

$$P\left(\left\|\bar{R} - \hat{Y}^T A \hat{Y}\right\|_F^2 \geq \psi\right) \leq \frac{p^2 \sigma_r^2}{L\psi}.$$

In particular, if $\sigma_a \Delta^{-1}$ does not vary with n, this theorem implies convergence of the angular phase with constant probability after $O(\epsilon^{-1} n p^3 \log n)$ iterations and in the same amount of time. Note that since we do not reuse samples in Alecton, our rates do not differentiate between sampling and computational complexity, unlike many other algorithms (see Appendix B). We also do not consider numerical error or overflow: periodically re-normalizing the iterate may be necessary to prevent these in an implementation of Alecton.

Since the upper bound expression uses Z_p, which is obscure, we plot it here (Figure 1). We also can make a more precise statement about the failure rate for $p = 1$.

Lemma 1. *For the case of rank-*1 *recovery,*

$$Z_1(\gamma) = \sqrt{2\pi\gamma} \exp\left(\frac{\gamma}{2}\right) \operatorname{erfc}\left(\sqrt{\frac{\gamma}{2}}\right) \leq \sqrt{2\pi\gamma}.$$

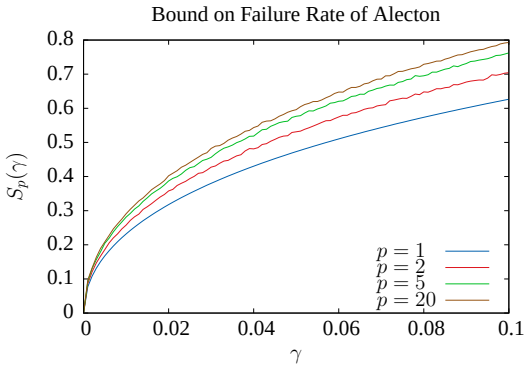


Figure 1: Value of Z_p computed as average of 10^5 samples.

3.1 Martingale Technique

A proof for Theorem 1 and full formal definitions will appear in Appendix C of this document, but since the method is nonstandard for non-convex optimization (although it has been used in Shamir [34] to show convergence for convex problems), we will outline it here. First, we define a *failure event* f_k at each timestep, that occurs if the iterate gets "too close" to the unstable fixed points. Next, we define a sequence τ_k, where

$$\tau_k = \frac{\left|Y_k^T U Y_k\right|}{\left|Y_k^T \left(\gamma n^{-1} p^{-2} q I + (1 - \gamma n^{-1} p^{-2} q) U\right) Y_k\right|}$$

(where $|X|$ denotes the determinant of X); the intuition here is that τ_k is close to 1 if and only if success occurs, and close to 0 when failure occurs. We show that, if neither success nor failure occurs at time k,

$$\mathbf{E}\left[\tau_{k+1} | \mathcal{F}_k\right] \geq \tau_k \left(1 + R\left(1 - \tau_k\right)\right) \tag{6}$$

for some constant R; here, $\mathcal{F}_k$ denotes the *filtration* at time k, which contains all the events that have occurred up to time k [17]. If we let T denote the first time at which either success or failure occurs, then this implies that τ_k is a submartingale for $k < T$. We use the optional stopping Theorem [17] (here we state a discrete-time version).

Definition 3 (Stopping Time)**.** A random variable T is a stopping time with respect to a filtration $\mathcal{F}_k$ if and only if $\{T \leq k\} \in \mathcal{F}_k$ for all k. That is, we can tell whether $T \leq k$ using only events that have occurred up to time k.

Theorem 2 (Optional Stopping Theorem)**.** *If x_k is a martingale (or submartingale) with respect to a filtration $\mathcal{F}_k$, and T is a stopping time with respect to the same filtration, then $x_{k\wedge T}$ is also a martingale (resp. submartingale) with respect to the same filtration, where $k \wedge T$ denotes the minimum of k and T. In particular, for bounded submartingales, this implies that $\mathbf{E}[x_0] \leq \mathbf{E}[x_T]$.*

Here, T is a stopping time since it depends only on events occurring before timestep T. Applying this to the submartingale τ_k results in

$$\begin{aligned}\mathbf{E}[\tau_0] &\leq \mathbf{E}[\tau_T] \\ &= \mathbf{E}[\tau_T|F_T]\,P(f_T) + \mathbf{E}[\tau_T|\neg F_T]\,(1 - P(f_T)) \\ &\leq \delta P(f_T) + (1 - P(f_T)).\end{aligned}$$

This isolates the probability of the failure event occurring. Next, subtracting 1 from both sides of (6) and taking the logarithm results in

$$\begin{aligned}\mathbf{E}[\log(1-\tau_{k+1})|\mathcal{F}_k] &\leq \log(1-\tau_k) + \log(1 - R\tau_k) \\ &\leq \log(1-\tau_k) - R\delta.\end{aligned}$$

So, if we let $W_k = \log(1-\tau_k) + R\delta k$, then W_k is a supermartingale. We again apply the optional stopping theorem to produce

$$\mathbf{E}[W_0] \geq \mathbf{E}[W_T] = \mathbf{E}[\log(1-\tau_T)] + R\delta\mathbf{E}[T].$$

This isolates the expected value of the stopping time. Finally, we notice that success occurs before time t if $T \leq t$ and f_T does not occur. By the union bound, this implies that

$$P_{\text{failure}} \leq P(f_T) + P(T \leq t),$$

and by Markov's inequality,

$$P_{\text{failure}} \leq P(f_T) + t^{-1}\mathbf{E}[T].$$

Substituting the isolated values for $P(f_T)$ and $\mathbf{E}[T]$ produces the expression above in (5).

The radial part of the theorem follows from an application of Chebychev's inequality to the average of L samples of $\hat{y}^T\tilde{A}\hat{y}$ — we do not devote any discussion to it since averages are already well understood.

4 Application Examples

4.1 Entrywise Sampling

One sampling distribution that arises in many applications (most importantly, matrix completion [12]) is *entrywise sampling*. This occurs when the samples are independently chosen from the entries of A. Specifically,

$$\tilde{A} = n^2 e_i e_i^T A e_j e_j^T,$$

where i and j are each independently drawn from $1, \ldots, n$. It is standard for these types of problems to introduce a *matrix coherence bound* [25].

Definition 4. A matrix $A \in \mathbb{R}^{n\times n}$ is incoherent with parameter μ if and only if for every unit eigenvector u_i of the matrix, and for all standard basis vectors e_j,

$$\left|e_j^T u_i\right| \leq \mu n^{-\frac{1}{2}}.$$

Under an incoherence assumption, we can provide a bound on the second moment of $\tilde{A}$, which is all that we need to apply Theorem 1 to this problem.

Lemma 2. *If A is incoherent with parameter μ, and $\tilde{A}$ is sampled uniformly from the entries of A, then the distribution of $\tilde{A}$ satisfies the Alecton variance condition with parameters $\sigma_a^2 = \mu^4 \|A\|_F^2$ and $\sigma_r^2 = \mu^4 \mathbf{tr}(A)^2$.*

For problems in which the matrix A is of constant rank, and its eigenvalues do not vary with n, neither $\|A\|_F$ nor $\mathbf{tr}(A)$ will vary with n. In this case, σ_a^2, σ_r^2, and Δ will be constants, and the $O(\epsilon^{-1} n \log n)$ bound on convergence time will hold.

4.2 Rectangular Entrywise Sampling

Entrywise sampling also commonly appear in rectangular matrix recovery problems. In these cases, we are trying to solve something like

$$\begin{array}{ll} \text{minimize} & \|M - X\|_F^2 \\ \text{subject to} & X \in \mathbb{R}^{m \times n}, \mathbf{rank}(X) \leq p. \end{array}$$

To solve this problem using Alecton, we first convert it into a symmetric matrix problem by constructing the block matrix

$$A = \begin{bmatrix} 0 & M \\ M^T & 0 \end{bmatrix};$$

it is known that recovering the dominant eigenvectors of A is equivalent to recovering the dominant singular vectors of M.

Entrywise sampling on M corresponds to choosing a random $i \in 1, \ldots, m$ and $j \in 1, \ldots, n$, and then sampling $\tilde{A}$ as

$$\tilde{A} = mnM_{ij}(e_i e_{m+j}^T + e_{m+j} e_i^T).$$

In the case where we can bound the entries of M (this is natural for recommender systems), we can prove the following.

Lemma 3. *If $M \in \mathbb{R}^{m \times n}$ satisfies the entry bound*

$$M_{ij}^2 \leq \xi m^{-1} n^{-1} \|M\|_F^2$$

for all i and j, then the rectangular entrywise sampling distribution on M satisfies the Alecton variance condition with parameters

$$\sigma_a^2 = \sigma_r^2 = 2\xi \|M\|_F^2.$$

As above, for problems in which the singular values of M do not vary with problem size, our big-O convergence time bound will still hold.

4.3 Trace Sampling

Another common sampling distribution arises from the *matrix sensing* problem [25]. In this problem, we are given the value of $v^T A w$ for unit vectors v and w selected uniformly at random. (This problem has been handled for the more general complex case in [11] using Wirtinger flow.) Using a trace sample, we can construct an unbiased sample

$$\tilde{A} = n^2 v v^T A w w^T.$$

This lets us bound the variance as follows.

Lemma 4. *If $n > 50$, and v and w are sampled uniformly from the unit sphere in $\mathbb{R}^n$, then for any positive semidefinite matrix A, if we let $\tilde{A} = n^2 vv^T Aww^T$, then the distribution of $\tilde{A}$ satisfies the Alecton variance condition with parameters $\sigma_a^2 = 16 \|A\|_F^2$ and $\sigma_r^2 = 16\mathbf{tr}\,(A)^2$.*

As above, for problems in which the eigenvalues of A do not vary with problem size, our big-O convergence time bound will still hold.

In some cases of the trace sampling problem, instead of being given samples of the form $u^T Av$, we know $u^T Au$. In this case, we need to use two independent samples $u_1^T Au_1$ and $u_2^T Au_2$, and let $u \propto u_1 + u_2$ and $v \propto u_1 - u_2$ be two unit vectors which we will use in the above sampling scheme. Notice that since u_1 and u_2 are independent and uniformly distributed, u and v will also be independent and uniformly distributed (by the spherical symmetry of the underlying distribution). Furthermore, we can compute

$$u^T Av = (u_1 + u_2)^T A(u_1 - u_2) = u_1^T Au_1 - u_2^T Au_2.$$

This allows us to use our above trace sampling scheme even with samples of the form $u^T Au$.

4.4 Subspace Sampling

Our analysis can handle more complicated sampling schemes. Consider the following distribution, which arises in subspace tracking [6]. Our matrix A is a rank-r projection matrix, and each sample consists of some randomly-selected entries from a randomly-selected vector in its column space. Specifically, we are given Qv and Rv, where v is some vector selected uniformly at random from $C(A)$, and Q and R are independent random diagonal projection matrices with expected value $mn^{-1}I$. Using this, we can construct the distribution

$$\tilde{A} = rn^2 m^{-2} Qvv^T R.$$

This distribution is unbiased since $\mathbf{E}\left[qvv^T\right] = A$. When bounding its second moment, we run into the same coherence problem as we did in the entrywise case, which motivates us to introduce a coherence constraint for subspaces.

Definition 5. A subspace of $\mathbb{R}^n$ of dimension q with associated projection matrix U is incoherent with parameter μ if and only if for all standard basis vectors e_i,

$$\|Ue_i\|^2 \leq \mu r n^{-1}.$$

Using this, we can prove the following facts about the second moment of this distribution.

Lemma 5. *The subspace sampling distribution, when sampled from a subspace that is incoherent with parameter μ, satisfies the Alecton variance condition with parameters*

$$\sigma_a^2 = \sigma_r^2 = r^2(1 + \mu r m^{-1})^2.$$

In many cases of subspace sampling, we are given just some entries of v at each timestep (as opposed to two separate random sets of entries associated with Q and R). That is, we are given a random diagonal projection matrix S, and the product Sv. We can use this to construct a sample of the above form by randomly splitting the given entries among Q and R in such a way that $Q = QS$ and $R = RS$, and Q and R are independent. We can then construct an unbiased sample as

$$\tilde{A} = rn^2 m^{-2} QSvv^T SR,$$

which uses only the entries of v that we are given.

4.5 Noisy Sampling

Since our analysis depends only on a variance bound, it is straightforward to handle the case in which the values of our samples themselves are noisy. Using the additive property of the variance for independent random variables, we can show that additive noise only increases the variance of the sampling distribution by a constant amount proportional to the variance of the noise. Similarly, using the multiplicative property of the variance for independent random variables, multiplicative noise only multiplies the variance of the sampling distribution by a constant factor proportional to the variance of the noise. In either case, we can show that the noisy sampling distribution satisfies AVC.

4.6 Extension to Higher Ranks

It is possible to use multiple iterations of the rank-1 version of Alecton to recover additional eigenvalue/eigenvector pairs of the data matrix A one-at-a-time. This is a standard technique for using power iteration algorithms to recover multiple eigenvalues. Sometimes, this may be preferable to using a single higher-rank invocation of Alecton (for example, we may not know a priori how many eigenvectors we want). We outline this technique as Algorithm 2. This strategy allows us to recover the largest p eigenvectors of A using p executions

Algorithm 2 Alecton One-at-a-time

Require: A sampling distribution $\mathcal{A}$
  $\mathcal{A}_1 \to \mathcal{A}$
  for $i = 1$ **to** p **do**
    ▷ Run rank-1 Alecton to produce output y_i.
    $y_i \to \text{Alecton}_{p=1}(\mathcal{A}_i)$
    Generate sampling distribution $\mathcal{A}_{i+1}$ such that, if $\tilde{A}'$ is sampled from $\mathcal{A}_{i+1}$ and $\tilde{A}$ is sampled from $\mathcal{A}_i$, $\mathbf{E}\left[\tilde{A}'\right] = \mathbf{E}\left[\tilde{A}\right] - y_i y_i^T$.
  end for
  return $\sum_{i=1}^{p} y_i y_i^T$

of Alecton. If the eigenvalues of the matrix are independent of n and p, we will be able to accomplish this in $O(\epsilon^{-1} pn \log n)$ total steps.

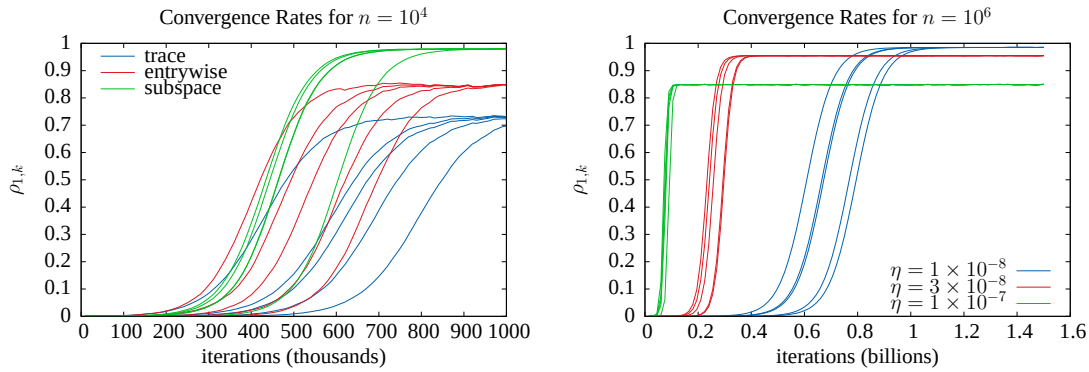
5 Experiments

We experimentally verify our main claim, that Alecton does converge quickly for practical datasets.

All experiments were run on a machine with a single twelve-core socket (Intel Xeon E5-2697, 2.70GHz), and 256 GB of shared memory. All were written in C++, excepting the Netflix Prize problem experiment, which was written in Julia. No data was collected for the radial phase of Alecton, since the performance of averaging is already well understood.

The first experiments were run on randomly-generated rank-10 data matrices $A \in \mathbb{R}^{n \times n}$. Each was generated by selecting a random orthogonal matrix $U \in \mathbb{R}^{n \times n}$, then independently selecting a diagonal matrix Λ with 10 positive nonzero eigenvalues, and constructing $A = U \Lambda U'$. Figure 2(a) illustrates the convergence of Alecton with $p = q = 1$ using three sampling distributions on datasets with $n = 10^4$. We ran Alecton starting from five random initial values; the different plotted trajectories illustrate how convergence time can depend on the initial value.

Figure 2(b) illustrates the performance of Alecton ($p = q = 1$ again) on a larger dataset with $n = 10^6$ as the step size parameter η is varied. As we would expect, a smaller value of η yields slower, but more



(a) Angular convergence of three distributions on a synthetic dataset with $\eta = 10^{-5}$.

(b) Angular convergence of entrywise sampling on a large synthetic dataset for different step sizes.

Figure 2: Convergence occurs in $O(n \log n)$ steps.

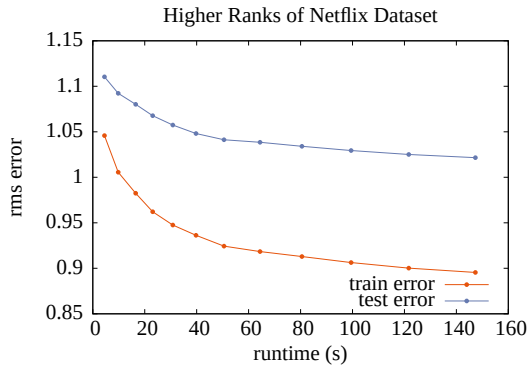


Figure 3: RMS errors over Netflix dataset [18] for higher-rank recovery. Each point represents an additional recovered eigenvector found with Alecton One-at-a-time.

accurate convergence. Also notice that the smaller the value of η, the more the initial value seems to affect convergence time.

Figure 3 demonstrates convergence results on real data from the Netflix Prize problem. This problem involves recovering a matrix with 480,189 columns and 17,770 rows from a training dataset containing 110,198,805 revealed entries. We used the rectangular entrywise distribution described above, then ran Alecton with $\eta = 10^{-12}$ and $p = q = 1$ for ten million iterations to recover the most significant singular vector. Next, we used Algorithm 2 to recover additional singular vectors of the matrix, up to a maximum of $p = 12$. The absolute runtime and RMS errors after the recovery of each subsequent eigenvector are plotted in Figure 3. This plot illustrates that the runtime of the one-at-a-time algorithm does not increase disastrously as the number of recovered eigenvectors expands.

5.1 Discussion

The Hogwild! algorithm [30] is a parallel, lock-free version of stochastic gradient descent that has been shown to perform similarly to sequential SGD on convex problems, while allowing for a good parallel speedup. It is an open question whether a Hogwild! version of Alecton for non-convex problems converges with a good rate, but we are optimistic that it will.

6 Conclusion

This paper exhibited Alecton, a stochastic gradient descent algorithm applied to a non-convex low-rank factorized problem; it is similar to the algorithms used in practice to solve a wide variety of problems. We prove that Alecton converges globally, and provide a rate of convergence. We do not require any special initialization step but rather initialize randomly. Furthermore, our result depends only on the variance of the samples, and therefore holds under broad sampling conditions that include both matrix completion and matrix sensing, and is also able to take noisy samples into account. We show these results using a martingale-based technique that is novel in the space of non-convex optimization, and we are optimistic that this technique can be applied to other problems in the future.

References

[1] P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, Princeton, NJ, 2008. ISBN 978-0-691-13298-3.

[2] Alekh Agarwal, Olivier Chapelle, Miroslav Dudík, and John Langford. A reliable effective terascale linear learning system. *CoRR*, abs/1110.4198, 2011.

[3] R. Arora, A. Cotter, K. Livescu, and N. Srebro. Stochastic optimization for pca and pls. In *Communication, Control, and Computing (Allerton), 2012 50th Annual Allerton Conference on*, pages 861–868, Oct 2012.

[4] Raman Arora, Andy Cotter, and Nati Srebro. Stochastic optimization of pca with capped msg. In C.j.c. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 1815–1823. 2013.

[5] Akshay Balsubramani, Sanjoy Dasgupta, and Yoav Freund. The fast convergence of incremental pca. In *NIPS*, pages 3174–3182, 2013.

[6] Laura Balzano, Robert Nowak, and Benjamin Recht. Online identification and tracking of subspaces from highly incomplete information. In *Communication, Control, and Computing (Allerton), 2010 48th Annual Allerton Conference on*, pages 704–711. IEEE, 2010.

[7] Lon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. 2010.

[8] Lon Bottou and Olivier Bousquet. The tradeoffs of large scale learning. In *IN: ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS 20*, pages 161–168, 2008.

[9] Samuel Burer and Renato DC Monteiro. A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. *Mathematical Programming*, 95(2):329–357, 2003.

[10] Samuel Burer and Renato DC Monteiro. Local minima and convergence in low-rank semidefinite programming. *Mathematical Programming*, 103(3):427–444, 2005.

[11] Emmanuel Candès, Xiaodong Li, and Mahdi Soltanolkotabi. Phase retrieval via wirtinger flow: Theory and algorithms. *arXiv preprint arXiv:1407.1065*, 2014.

[12] Emmanuel J. Candès and Benjamin Recht. Exact matrix completion via convex optimization. *FoCM*, 9(6):717–772, 2009. ISSN 1615-3375.

[13] EmmanuelJ. Candès and Xiaodong Li. Solving quadratic equations via phaselift when there are about as many equations as unknowns. *FoCM*, 14(5):1017–1026, 2014.

[14] Caihua Chen, Bingsheng He, and Xiaoming Yuan. Matrix completion via an alternating direction method. *IMAJNA*, 2011.

[15] M.P. do Carmo. *Riemannian Geometry*. Mathematics (Birkhäuser) theory. Birkhäuser Boston, 1992. ISBN 9780817634902.

[16] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *J. Mach. Learn. Res.*, 12:2121–2159, July 2011.

[17] Thomas R Fleming and David P Harrington. Counting processes and survival analysis. volume 169, pages 56–57. John Wiley & Sons, 1991.

[18] Simon Funk. Netflix Update: Try this at Home. 2006.

[19] Michel X. Goemans and David P. Williamson. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *J. ACM*, 42(6):1115–1145, November 1995.

[20] Pankaj Gupta, Ashish Goel, Jimmy Lin, Aneesh Sharma, Dong Wang, and Reza Zadeh. Wtf: The who to follow service at twitter. WWW '13, pages 505–514, 2013.

[21] Moritz Hardt and Eric Price. The noisy power method: A meta algorithm with applications. In Z. Ghahramani, M. Welling, C. Cortes, N.d. Lawrence, and K.q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2861–2869. Curran Associates, Inc., 2014.

[22] Steven Homer and Marcus Peinado. Design and performance of parallel and distributed approximation algorithms for maxcut. *J. Parallel Distrib. Comput.*, 46(1):48–61, October 1997. ISSN 0743-7315.

[23] R. Horstmeyer, R. Y. Chen, X. Ou, B. Ames, J. A. Tropp, and C. Yang. Solving ptychography with a convex relaxation. *ArXiv e-prints*, dec 2014.

[24] Chonghai Hu, James T. Kwok, and Weike Pan. Accelerated gradient methods for stochastic optimization and online learning. In *Advances in Neural Information Processing Systems 22*, pages 781–789, 2009.

[25] Prateek Jain, Praneeth Netrapalli, and Sujay Sanghavi. Low-rank matrix completion using alternating minimization. In *Proceedings of the Forty-fifth Annual ACM STOC*, pages 665–674. ACM, 2013.

[26] Raman Arora John Goes, Teng Zhang and Gilad Lerman. Robust stochastic principal component analysis. In *Proceedings of the 17th International Conference on Artificial Intelligence and Statistics*, pages 266–274, 2014.

[27] M. Journée, F. Bach, P.-A. Absil, and R. Sepulchre. Low-rank optimization on the cone of positive semidefinite matrices. *SIAM J. on Optimization*, 20(5):2327–2351, May 2010.

[28] R.H. Keshavan, A. Montanari, and Sewoong Oh. Matrix completion from a few entries. *Information Theory, IEEE Transactions on*, 56(6):2980–2998, June 2010.

[29] Bamdev Mishra, Gilles Meyer, Francis Bach, and Rodolphe Sepulchre. Low-rank optimization with trace norm penalty. *SIAM Journal on Optimization*, 23(4):2124–2149, 2013.

[30] Feng Niu, Benjamin Recht, Christopher Ré, and Stephen J. Wright. Hogwild: A lock-free approach to parallelizing stochastic gradient descent. In *In NIPS*, 2011.

[31] Erkki Oja. On stochastic approximation of the eigenvectors and eigenvalues of the expectation of a random matrix. *Journal of Mathematical Analysis and Applications*, 106, 1985.

[32] Ian O'Connell Jimmy Lin Oscar Boykin, Sam Ritchie. Summingbird: A framework for integrating batch and online mapreduce computations. In *Proceedings of the VLDB Endowment*, volume 7, pages 1441–1451, 2013-2014.

[33] Benjamin Recht and Christopher Ré. Parallel stochastic gradient algorithms for large-scale matrix completion. *Mathematical Programming Computation*, 5(2):201–226, 2013. ISSN 1867-2949.

[34] Ohad Shamir. Making gradient descent optimal for strongly convex stochastic optimization. *CoRR*, abs/1109.5647, 2011.

[35] Ohad Shamir. A stochastic PCA algorithm with an exponential convergence rate. *CoRR*, abs/1409.2848, 2014.

[36] Christina Teflioudi, Faraz Makari, and Rainer Gemulla. Distributed matrix completion. *2013 IEEE 13th ICDM*, 0:655–664, 2012. ISSN 1550-4786.

[37] Hui Zou, Trevor Hastie, and Robert Tibshirani. Sparse principal component analysis. *J. Comp. Graph. Stat.*, 15:2006, 2004.

A Negative Results

Divergence Example Here, we observe what happens when we choose a constant step size for stochastic gradient descent for quartic objective functions. Consider the simple optimization problem of minimizing

$$f(x) = \frac{1}{4}x^4.$$

This function will have gradient descent update rule

$$x_{k+1} = x_k - \alpha_k x_k^3 = \left(1 - \alpha_k x_k^2\right) x_k.$$

We now prove that, for any reasonable step size rule chosen independently of x_k, there is some initial condition such that this iteration diverges to infinity.

Proposition 1. *Assume that we iterate using the above rule, for some choice of α_k that is not super-exponentially decreasing; that is, for some $C > 1$ and some $\alpha > 0$, $\alpha_k \geq \alpha C^{-2k}$ for all k. Then, if $x_0^2 \geq \alpha^{-1}(C+1)$, for all k*

$$x_k^2 > \alpha^{-1} C^{2k} (C+1).$$

Proof. We will prove this by induction. The base case follows directly from the assumption, while under the inductive case, if the proposition is true for k, then

$$\alpha_k x_k^2 \geq \alpha C^{-2k} \alpha^{-1} C^{2k} (C+1) = C+1.$$

Therefore,

$$\begin{aligned} x_{k+1}^2 &= \left(\alpha_k x_k^2 - 1\right)^2 x_k^2 \\ &\geq C^2 x_k^2 \\ &\geq C^2 \alpha^{-1} C^{2k} (C+1) \\ &= \alpha^{-1} C^{2(k+1)} (C+1). \end{aligned}$$

This proves the statement. □

This proof shows that, for some choice of x_0, x_k will diverge to infinity exponentially quickly. Furthermore, no reasonable choice of α_k will be able to halt this increase for all initial conditions. We can see the effect of this in stochastic gradient descent as well, where there is always some probability that, due to an unfortunate series of gradient steps, we will enter the zone in which divergence occurs. On the other hand, if we chose step size $\alpha_k = \gamma_k x_k^{-2}$, for some $0 < \gamma_k < 2$, then

$$x_{k+1} = (1 - \gamma_k)\, x_k,$$

which converges for all starting values of x_k. This simple example is what motivates us to take $\|Y_k\|$ into account when choosing the step size for Alecton.

Global Convergence Counterexample We now exhibit a particular problem for which SGD on a low-rank factorization doesn't converge to the global optimum for a particular starting point. Let matrix $A \in \mathbb{R}^{2\times 2}$ be the diagonal matrix with diagonal entries 4 and 1. Further, let's assume that we are trying to minimize the expected value of the decomposed rank-1 objective function

$$\tilde{f}(y) = \left\| \tilde{A} - yy^T \right\|_F = \|y\|^4 - 2y^T \tilde{A} y + \left\| \tilde{A} \right\|_F^2.$$

If our stochastic samples satisfy $\tilde{A} = A$ (i.e. we use a perfect sampler), then the SGD update rule is

$$y_{k+1} = y_k - \alpha_k \nabla \tilde{f}(y_k) = y_k - 4\alpha_k \left(y_k \|y_k\|^2 - A y_k \right).$$

Now, we know that e_1 is the most significant eigenvector of A, and that $y = 2e_1$ is the global solution to the problem. However,

$$\begin{aligned} e_1^T y_{k+1} &= e_1^T y_k - 4\alpha_k \left(e_1^T y_k \|y_k\|^2 - e_1^T A y_k \right) \\ &= \left(1 - 4\alpha_k \left(\|y_k\|^2 - 4 \right) \right) e_1^T y_k \end{aligned}$$

. This implies that if $e_1^T y_0 = 0$, then $e_1^T y_k = 0$ for all k, which means that convergence to the global optimum cannot occur. This illustrates that global convergence does not occur for all manifold optimization problems using a low-rank factorization and for all starting points.

Constraints Counterexample We might think that our results can be generalized to give $O(n \log n)$ convergence of low-rank factorized problems with arbitrary constraints. Here, we show that this will not work for all problems by encoding an NP-complete problem as a constrained low-rank optimization problem.

For any graph with node set N and edge set E, the MAXCUT problem on the graph requires us to solve

$$\begin{array}{ll} \text{minimize} & \sum_{(i,j)\in E} y_i y_j \\ \text{subject to} & y_i \in \{-1, 1\}. \end{array}$$

Algorithm	Sampling Scheme	Complexity	
		Sampling	Computational
Alecton	Any	$O(\epsilon^{-1}p^3 n \log n)$	
SVD	Various	$o(pn)$	$O(n^3)$
Spectral Matrix Completion [28]	Elementwise	$o(pn)$	$O(p^2 n \log n)$
PhaseLift [13]	Phase Retrieval	$o(n)$	$O(\epsilon^{-1}n^3)$
Alternating Minimization [41]	Phase Retrieval	$o(n \log(\epsilon^{-1}))$	$O(n^2 \log^2(\epsilon^{-1}))$
Wirtinger Flow [11]	Phase Retrieval	$o(n \log^2 n)$	$O(pn \log(\epsilon^{-1}))$

Equivalently, if we let A denote the edge-matrix of the graph, we can represent this as a matrix problem [19, 22]

$$\begin{array}{ll} \text{minimize} & y^T A y \\ \text{subject to} & y_i \in \{-1, 1\}. \end{array}$$

We relax this problem to

$$\begin{array}{ll} \text{minimize} & y^T A y \\ \text{subject to} & -1 \leq y_i \leq 1. \end{array}$$

Since the diagonal of A is zero, if we fix all but one of the entries of y, the objective function will have an affine dependence on that entry. In particular, this means that a global minimum of the problem must occur on the boundary where $y_i \in \{-1, 1\}$, which implies that this problem has the same global solution as the original MAXCUT problem. Furthermore, for sufficiently large values of σ, the problem

$$\begin{array}{ll} \text{minimize} & \|y\|^4 + 2\sigma y^T A y + \sigma^2 \|A\|_F^2 \\ \text{subject to} & -1 \leq y_i \leq 1 \end{array}$$

will also have the same solution. But, this problem is in the same form as a low-rank factorization of

$$\begin{array}{ll} \text{minimize} & \|X + \sigma A\|_F^2 \\ \text{subject to} & X_{ii} \leq 1, X \succeq 0, \mathbf{rank}\,(X) = 1 \end{array}$$

where $X = yy^T$. Since MAXCUT is NP-complete, it can't possibly be the case that SGD applied to this low-rank factorized problem converges quickly to the global optimum, because that would imply an efficient solution to this NP-complete problem. This suggests that care will be needed when analyzing problems with constraints, in order to exclude these sorts of cases.

B Comparison with Other Methods

There are several other algorithms that solve similar matrix recover problems in the literature. In Table B, we list some other algorithms, and their convergence rates, in terms of both number of samples required (sampling complexity) and number of iterations performed (computational complexity). For this table, the data is assumed to be of dimension n, and the rank (where applicable) is assumed to be p. (In order to save space, factors of $\log \log \epsilon^{-1}$ have been omitted from some formulas.)

C Proofs of Main Results

In this appendix, we provide rigorous definitions and detail the proof outlined in Section 3.1.

C.1 Definitions

Fleming and Harrington [17] provide the following definitions of filtration and martingale. We state the definitions adapted to the discrete-time case.

Definition 6 (Filtration)**.** Given a measurable probability space $(\Omega, \mathcal{F})$, a *filtration* is a sequence of sub-σ-algebras $\{\mathcal{F}_t\}$ for $t \geq 0$, such that for all $s \leq t$,

$$\mathcal{F}_s \subset \mathcal{F}_t.$$

That is, if an event A is in $\mathcal{F}_s$, and $t \geq s$, then A is also in $\mathcal{F}_t$. This definition encodes the monotonic increase in available information over time.

Definition 7 (Martingale)**.** Let $\{X_t\}$ be a stochastic process and $\{\mathcal{F}_t\}$ be a filtration over the same probability space. Then X is called a *martingale* with respect to the filtration if for every t, X_t is $\mathcal{F}_t$-measurable, and

$$\mathbf{E}\left[X_{t+1}|\mathcal{F}_t\right] = X_t. \tag{7}$$

We call X a *submartingale* if the same conditions hold, except (7) is replaced with

$$\mathbf{E}\left[X_{t+1}|\mathcal{F}_t\right] \geq X_t.$$

We call X a *supermartingale* if the same conditions hold, except (7) is replaced with

$$\mathbf{E}\left[X_{t+1}|\mathcal{F}_t\right] \leq X_t.$$

C.2 Preliminaries

In addition to the quantities used in the statement of Theorem 1, we let

$$W = \gamma n^{-1} p^{-2} q I + (1 - \gamma n^{-1} p^{-2} q) U,$$

and define sequences τ_k and ϕ_k as

$$\tau_k = \frac{\left|Y_k^T U Y_k\right|}{\left|Y_k^T W Y_k\right|},$$

and

$$\phi_k = \mathbf{tr}\left(I - Y_k^T U Y_k \left(Y_k^T W Y_k\right)^{-1}\right).$$

This agrees with the definition of τ_k stated in the body of the paper. Using this sequence, we define the failure event f_k as the event that occurs when

$$\tau_k \leq \frac{1}{2}. \tag{8}$$

We recall that we defined the success event at time k as the event that, for all $z \in \mathbb{R}^p$,

$$\frac{\left\|U Y_k z\right\|^2}{\left\|Y_k z\right\|^2} \geq 1 - \epsilon.$$

Finally, we define T, the stopping time, to be the first time at which either the success event or the failure event occurs.

Now, we state some lemmas we will need in the following proofs. We defer proofs of the lemmas themselves to Appendix D. First, we state a lemma about quadratic rational functions that we will need in the next section.

Lemma 6 (Quadratic rational lower bound)**.** *For any a, b, c, and d in $\mathbb{R}$, if $1+by+cy^2 > 0$ and $1+ay+dy^2 \geq 0$ for all y, then for all $x \in \mathbb{R}$,*

$$\frac{1+ax+dx^2}{1+bx+cx^2} \geq 1+(a-b)x-cx^2.$$

Next, a lemma about the expected initial value of τ:

Lemma 7. *If we initialize Y_0 uniformly as in the Alecton algorithm, then*

$$\mathbf{E}\left[\tau_0\right] \geq 1 - \frac{1}{2}Z_p(\gamma).$$

Next, a lemmas that bounds a determinant expression.

Lemma 8. *For any $B \in \mathbb{R}^{n\times n}$, $Y \in \mathbb{R}^{n\times m}$, and any symmetric positive-semidefinite $Z \in \mathbb{R}^{n\times n}$, if either B is rank-1 or $m = 1$, then*

$$\begin{aligned}&\left|Y^T(I+B)^TZ(I+B)Y\right| \\ &\quad\geq \left|Y^TZY\right|\left(\mathbf{tr}\left(Y(Y^TZY)^{-1}Y^TZB\right)+1\right)^2\end{aligned}$$

and

$$\begin{aligned}&\left|Y^T(I+B)^TZ(I+B)Y\right| \\ &\quad\leq \left|Y^TZY\right|\Big(1+2\mathbf{tr}\left(Y(Y^TZY)^{-1}Y^TZB\right) \\ &\qquad+\mathbf{tr}\left(Y(Y^TZY)^{-1}Y^TB^TZB\right)\Big).\end{aligned}$$

Next, a lemma that bounds τ in the case that the success condition does not occur.

Lemma 9. *If we run Alecton, and at timestep k, the success condition does not hold, then*

$$\tau_k \leq 1 - \gamma n^{-1}p^{-2}q\epsilon.$$

Finally, a lemma that relates ϕ and τ.

Lemma 10. *Using the definitions above, for all k,*

$$\phi_k \geq 1 - \tau_k.$$

C.3 Main Proofs

We now proceed to prove Theorem 1 in six steps, as outlined in Section 3.1.

- First, we prove Lemma 11, the *dominant mass bound lemma*, which bounds $\mathbf{E}\left[\tau_{k+1}|\mathcal{F}_k\right]$ from below by a quadratic function of the step size η.
- We use this to prove Lemma 12, which establishes the result stated in (6).
- We use the optional stopping theorem to prove Lemma 13, which bounds the probability of a failure event occurring before success.
- We use the optional stopping theorem again to prove Lemma 14, which bounds the expected time until either a failure or success event occurs.

- We use Markov's inequality and the union bound to bound the angular failure probability of Theorem 1.
- Finally, we prove the radial phase result stated in Theorem 1.

Lemma 11 (Dominant Mass Bound)**.** *If we run Alecton under the conditions of Theorem 1, then for any k,*

$$\begin{aligned}\mathbf{E}\left[\tau_{k+1}|\mathcal{F}_k\right] \geq \tau_k\Big(1 + 2\eta\left(\Delta - \eta\sigma_a^2\gamma^{-1}np^2\right)(1-\tau_k) \\ - \eta^2\sigma_a^2 p(q+1)\Big).\end{aligned}$$

Proof. From the definition of τ, at the next timestep we will have

$$\begin{aligned}\tau_{k+1} &= \frac{\left|Y_{k+1}^T U Y_{k+1}\right|}{\left|Y_{k+1}^T W Y_{k+1}\right|} \\ &= \frac{\left|Y_k^T\left(I+\eta\tilde{A}_k\right)^T U\left(I+\eta\tilde{A}_k\right)Y_k\right|}{\left|Y_k^T\left(I+\eta\tilde{A}_k\right)^T W\left(I+\eta\tilde{A}_k\right)Y_k\right|}.\end{aligned}$$

Now, since our instance of Alecton satisfies the rank condition, either $\tilde{A}_k$ is rank-1 or $p = 1$. Therefore, we can apply Lemma 8 to these determinant quantities. In order to produce a lower bound on τ_{k+1}, we will apply lower bound to the numerator and the upper bound to the denominator. If we let $B_k = Y_k(Y_k^T U Y_k)^{-1}Y_k^T$, and $C_k = Y_k(Y_k^T W Y_k)^{-1}Y_k^T$, then this results in

$$\begin{aligned}\tau_{k+1} \geq \frac{\left|Y_k^T U Y_k\right|}{\left|Y_k^T W Y_k\right|} \\ \cdot \frac{\left(1+\eta\mathbf{tr}\left(B_k U\tilde{A}_k\right)\right)^2}{1+2\eta\mathbf{tr}\left(C_k W\tilde{A}_k\right)+\eta^2\mathbf{tr}\left(C_k\tilde{A}_k^T W\tilde{A}_k\right)}.\end{aligned}$$

Next, we apply Lemma 6, which results in

$$\begin{aligned}\tau_{k+1} &\geq \tau_k\Big(1+2\eta\left(\mathbf{tr}\left(B_k U\tilde{A}_k\right)-\mathbf{tr}\left(C_k W\tilde{A}_k\right)\right) \\ &\qquad -\eta^2\mathbf{tr}\left(C_k\tilde{A}_k^T W\tilde{A}_k\right)\Big) \\ &\geq \tau_k\left(1+2\eta R_k+\eta^2 Q_k\right),\end{aligned}$$

for sequences R_k and Q_k. Now, we investigate the expected values of these sequences. First, since the estimator has $\mathbf{E}\left[\tilde{A}_k\middle|\mathcal{F}_k\right] = A$, the expected value of R_k is

$$\begin{aligned}\mathbf{E}\left[R_k|\mathcal{F}_k\right] &= \mathbf{tr}\left(B_k U A\right) - \mathbf{tr}\left(C_k W A\right) \\ &= \mathbf{tr}\left((B_k - C_k)UA\right) \\ &\qquad -\gamma n^{-1}p^{-2}q\mathbf{tr}\left(C_k(I-U)A\right).\end{aligned}$$

Now, since U commutes with A, we will have that

$$UA \succeq \lambda_q U,$$

and similarly

$$(I-U)A \preceq \lambda_{q+1}(I-U).$$

Applying this results in

$$\begin{aligned}\mathbf{E}\left[R_k|\mathcal{F}_k\right] &\geq \mathbf{tr}\left(B_kUA\right) - \mathbf{tr}\left(C_kWA\right)\\ &= \lambda_q\mathbf{tr}\left((B_k-C_k)U\right)\\ &\quad - \lambda_{q+1}\gamma n^{-1}p^{-2}q\mathbf{tr}\left(C_k(I-U)\right).\end{aligned}$$

Now, we first notice that

$$\begin{aligned}\mathbf{tr}\left((B_k-C_k)U\right) &= \mathbf{tr}\left(I - Y_kUY_k^T(Y_k^TWY_k)^{-1}\right)\\ &= \phi_k.\end{aligned}$$

We also notice that

$$\begin{aligned}\gamma n^{-1}p^{-2}q\mathbf{tr}\left(C_k(I-U)\right) &= \mathbf{tr}\left(C_k(W-U)\right)\\ &= \mathbf{tr}\left(I - Y_kUY_k^T(Y_k^TWY_k)^{-1}\right)\\ &= \phi_k.\end{aligned}$$

It therefore follows that

$$\begin{aligned}\mathbf{E}\left[R_k|\mathcal{F}_k\right] &\geq (\lambda_q - \lambda_{q+1})\phi_k\\ &= \Delta\phi_k.\end{aligned}$$

Next, the expected value of Q_k is

$$\mathbf{E}\left[Q_k|\mathcal{F}_k\right] = \mathbf{tr}\left(C_k\mathbf{E}\left[\tilde{A}_k^TW\tilde{A}_k\right]\right).$$

Since our instance of Alecton satisfies the variance condition, and W commutes with A,

$$\mathbf{E}\left[Q_k|\mathcal{F}_k\right] \leq \sigma_a^2\mathbf{tr}\left(W\right)\mathbf{tr}\left(C_k\right).$$

We notice that

$$\begin{aligned}\mathbf{tr}\left(C_k\right) &= \mathbf{tr}\left(C_k\left(W + (1-\gamma n^{-1}p^{-2}q)(I-U)\right)\right)\\ &= p + (1-\gamma n^{-1}p^{-2}q)\mathbf{tr}\left(C_k(I-U)\right)\\ &\leq p + \mathbf{tr}\left(C_k(I-U)\right).\end{aligned}$$

By the logic above,

$$\mathbf{tr}\left(C_k\right) \leq p + \gamma^{-1}np^2q^{-1}\phi_k.$$

Also,

$$\begin{aligned}\mathbf{tr}\left(W\right) &= \mathbf{tr}\left(\gamma n^{-1}p^{-2}qI + (1-\gamma n^{-1}p^{-2}q)U\right)\\ &= \gamma p^{-2}q + q - \gamma n^{-1}p^{-2}q^2\\ &\geq q+1\end{aligned}$$

and therefore, since $\mathbf{tr}\,(W) \leq q+1$,

$$\mathbf{E}\left[Q_k|\mathcal{F}_k\right] \leq \sigma_a^2(q+1)\left(p+\gamma^{-1}np^2q^{-1}\phi_k\right).$$

Substituting these in results in

$$\begin{aligned}\mathbf{E}\left[\tau_{k+1}|\mathcal{F}_k\right] &\geq \tau_k\left(1+2\eta\Delta\phi_k-\eta^2\left(\sigma_a^2p(q+1)+\sigma_a^2\gamma^{-1}np^2(q+1)q^{-1}\phi_k\right)\right)\\ &= \tau_k\left(1+\eta\left(2\Delta-\eta\sigma_a^2\gamma^{-1}np^2(q+1)q^{-1}\right)\phi_k-\eta^2\sigma_a^2p(q+1)\right)\\ &\geq \tau_k\left(1+2\eta\left(\Delta-\eta\sigma_a^2\gamma^{-1}np^2\right)\phi_k-\eta^2\sigma_a^2p(q+1)\right).\end{aligned}$$

Finally, since for our chosen value of γ,

$$\Delta > \eta\sigma_a^2\gamma^{-1}np^2,$$

we can apply Lemma 10, which produces

$$\begin{aligned}\mathbf{E}\left[\tau_{k+1}|\mathcal{F}_k\right] \geq \tau_k\Big(&1+2\eta\left(\Delta-\eta\sigma_a^2\gamma^{-1}np^2\right)(1-\tau_k)\\ &-\eta^2\sigma_a^2p(q+1)\Big).\end{aligned}$$

This is the desired expression. □

Lemma 12. *If we run Alecton under the conditions of Theorem 1, then for any time k at which neither the success event nor the failure event occur,*

$$\mathbf{E}\left[\tau_{k+1}|\mathcal{F}_k\right] \geq \tau_k\left(1+\eta\Delta(1-\tau_k)\right).$$

Proof. From the result of Lemma 11,

$$\begin{aligned}\mathbf{E}\left[\tau_{k+1}|\mathcal{F}_k\right] &\geq \tau_k\left(1+2\eta\left(\Delta-\eta\sigma_a^2\gamma^{-1}np^2\right)(1-\tau_k)-\eta^2\sigma_a^2p(q+1)\right)\\ &= \tau_k\left(1+\eta\Delta(1-\tau_k)+\eta\left(\Delta-2\eta\sigma_a^2\gamma^{-1}np^2\right)(1-\tau_k)-\eta^2\sigma_a^2p(q+1)\right)\\ &= \tau_k\left(1+\eta\Delta(1-\tau_k)+\eta S_k,\right)\end{aligned}$$

for sequence S_k. Now, it can be easily verified that we chose γ such that

$$\Delta \geq 2\eta\sigma_a^2\gamma^{-1}np^2,$$

and so it follows that, by Lemma 9,

$$\begin{aligned}S_k &= \left(\Delta-2\eta\sigma_a^2\gamma^{-1}np^2\right)(1-\tau_k)-\eta\sigma_a^2p(q+1)\\ &\geq \left(\Delta-2\eta\sigma_a^2\gamma^{-1}np^2\right)\gamma n^{-1}p^{-2}q\epsilon-\eta\sigma_a^2p(q+1)\\ &= \Delta\gamma n^{-1}p^{-2}q\epsilon-2\eta\sigma_a^2q\epsilon-\eta\sigma_a^2p(q+1)\\ &\geq \Delta\gamma n^{-1}p^{-2}q\epsilon-2\eta\sigma_a^2q(p+\epsilon).\end{aligned}$$

If we substitute the value of γ,

$$\gamma = \frac{2n\sigma_a^2p^2(p+\epsilon)}{\Delta\epsilon}\eta.$$

then we arrive at

$$S_k \geq 0.$$

Substituting this in to our original expression produces

$$\mathbf{E}\left[\tau_{k+1}|\mathcal{F}_k\right] \geq \tau_k\left(1+\eta\Delta(1-\tau_k)\right),$$

as desired. □

Lemma 13 (Failure Probability Bound)**.** *If we run Alecton under the conditions of Theorem 1, then the probability that the failure event will occur before the success event is*

$$P\left(f_T\right) \leq Z_p(\gamma).$$

Proof. To prove this, we use the stopping time T, which we defined as the first time at which either the success event or failure event occurs. First, if $k < T$, it follows that neither success nor failure have occurred yet, so we can apply Lemma 12, which results in

$$\mathbf{E}\left[\tau_{k+1}|\mathcal{F}_k\right] \geq \tau_k\left(1+\eta\Delta(1-\tau_k)\right).$$

Therefore τ_k is a supermartingale for $k < T$. So, we can apply the optional stopping theorem, which produces

$$\mathbf{E}\left[\tau_0\right] \leq \mathbf{E}\left[\tau_T\right].$$

So, by the law of total expectation,

$$\mathbf{E}\left[\tau_0\right] \leq \mathbf{E}\left[\tau_T|f_T\right]P\left(f_T\right) + \mathbf{E}\left[\tau_T|\neg f_T\right]P\left(\neg f_T\right),$$

where f_T is the failure event at time T. Applying the definition of the failure event from (8),

$$\mathbf{E}\left[\tau_0\right] \leq \frac{1}{2}P\left(f_T\right) + 1\left(1 - P\left(f_T\right)\right).$$

Therefore, solving for $P\left(f_T\right)$,

$$P\left(f_T\right) \leq 2\left(1 - \mathbf{E}\left[\tau_0\right]\right).$$

Now applying Lemma 7,

$$P\left(f_T\right) \leq 2\left(1 - \left(1 - \frac{1}{2}Z_p(\gamma)\right)\right) = Z_p(\gamma),$$

as desired. □

Lemma 14 (Stopping Time Expectation)**.** *If we run Alecton under the conditions of Theorem 1, then the expected value of the stopping time T will be*

$$\mathbf{E}\left[T\right] \leq \frac{4n\sigma_a^2p^2(p+\epsilon)}{\Delta^2\gamma\epsilon}\log\left(\frac{np^2}{\gamma q\epsilon}\right).$$

Proof. First, as above if $k < T$, we can apply Lemma 12, which results in

$$\begin{aligned}\mathbf{E}\left[\tau_{k+1}|\mathcal{F}_k\right] &\geq \tau_k\left(1+\eta\Delta\left(1-\tau_k\right)\right)\\ &= \tau_k + \eta\Delta\tau_k\left(1-\tau_k\right),\end{aligned}$$

and so

$$\mathbf{E}\left[1-\tau_{k+1}|\mathcal{F}_k\right] \leq \left(1-\tau_k\right)\left(1-\eta\Delta\tau_k\right).$$

Now, if $k < T$, then since failure hasn't occurred yet, $\tau_k > \frac{1}{2}$. So,

$$\mathbf{E}\left[1-\tau_{k+1}|\mathcal{F}_k\right] \leq \left(1-\tau_k\right)\left(1-\frac{1}{2}\eta\Delta\right).$$

Now, since the logarithm function is concave, by Jensen's inequality we have

$$\mathbf{E}\left[\log\left(1-\tau_{k+1}\right)|\mathcal{F}_k\right] \leq \log\mathbf{E}\left[1-\tau_{k+1}|\mathcal{F}_k\right],$$

and thus by transitivity,

$$\begin{aligned}\mathbf{E}\left[\log\left(1-\tau_{k+1}\right)|\mathcal{F}_k\right] &\leq \log(1-\tau_k) + \log\left(1-\frac{1}{2}\eta\Delta\right) \\ &\leq \log(1-\tau_k) - \frac{1}{2}\eta\Delta.\end{aligned}$$

Now, we define a new process ψ_k as

$$\psi_k = \log(1-\tau_k) + \frac{1}{2}\eta\Delta k.$$

Using this definition, for $k < T$,

$$\begin{aligned}\mathbf{E}\left[\psi_{k+1}|\mathcal{F}_k\right] &= \mathbf{E}\left[\log(1-\tau_{k+1})|\mathcal{F}_k\right] + \frac{1}{2}\eta\Delta(k+1) \\ &\leq \log(1-\tau_k) - \frac{1}{2}\eta\Delta + \frac{1}{2}\eta\Delta(k+1) \\ &= \log(1-\tau_k) + \frac{1}{2}\eta\Delta k \\ &= \psi_k,\end{aligned}$$

so ψ_k is a supermartingale for $k < T$. We can therefore apply the optional stopping theorem, which states that

$$\mathbf{E}\left[\log(1-\tau_0)\right] = \mathbf{E}\left[\psi_0\right] \geq \mathbf{E}\left[\psi_T\right].$$

Since $1-\tau_0 < 1$, it follows that $\log(1-\tau_0) < 0$. Therefore,

$$0 \geq \mathbf{E}\left[\psi_T\right] = \mathbf{E}\left[\log(1-\tau_T)\right] + \frac{1}{2}\eta\Delta\mathbf{E}\left[T\right].$$

Applying Lemma 9,

$$1-\tau_T \geq \gamma n^{-1}p^{-2}q\epsilon,$$

and so

$$0 \geq \log(\gamma n^{-1}p^{-2}q\epsilon) + \frac{1}{2}\eta\Delta\mathbf{E}\left[T\right].$$

Solving for the expected value of the stopping time,

$$\mathbf{E}\left[T\right] \leq \frac{2}{\eta\Delta\delta}\log\left(\frac{np^2}{\gamma q\epsilon}\right).$$

Finally, substituting η in terms of γ results in

$$\mathbf{E}\left[T\right] \leq \frac{4n\sigma_a^2p^2(p+\epsilon)}{\Delta^2\gamma\epsilon}\log\left(\frac{np^2}{\gamma q\epsilon}\right),$$

as desired. □

Finally, we prove Theorem 1.

Proof of angular part of Theorem 1. First, we notice that the total failure event up to time t can be written as

$$F_t = f_T \cup \{T > t\}.$$

That is, total failure up to time t occurs if either failure happens before success (event f_T), or neither success nor failure happen before t. By the union bound,

$$F_t \leq P\left(f_T\right) + P\left(T > t\right).$$

Applying Markov's inequality,

$$P\left(F_t\right) \leq P\left(f_T\right) + \frac{1}{t}\mathbf{E}\left[T\right].$$

Finally, applying Lemmas 13 and 14 produces

$$P\left(F_t\right) \leq Z_p(\gamma) + \frac{4n\sigma_a^2 p^2(p+\epsilon)}{\Delta^2 \gamma \epsilon t} \log\left(\frac{np^2}{\gamma q \epsilon}\right).$$

This is the desired expression. □

Proof of radial part of Theorem 1. Recall that in Alecton, $\bar{R}$ is defined as

$$\bar{R} = \frac{1}{L}\sum_{l=0}^{L-1} \hat{Y}^T \tilde{A}_l \hat{Y}.$$

Now, computing the expected distance to the mean,

$$\begin{aligned}
&\mathbf{E}\left[\left\|\bar{R} - \hat{Y}^T A \hat{Y}\right\|_F^2\right] \\
&= \mathbf{E}\left[\left\|\frac{1}{L}\sum_{l=0}^{L-1} \hat{Y}^T \tilde{A}_l \hat{Y} - \hat{Y}^T A \hat{Y}\right\|_F^2\right] \\
&= \mathbf{E}\left[\left\|\frac{1}{L}\sum_{l=0}^{L-1} \hat{Y}^T (\tilde{A}_l - A) \hat{Y}\right\|_F^2\right] \\
&= \frac{1}{L^2}\mathbf{E}\left[\sum_{k=0}^{L-1}\sum_{l=0}^{L-1} \mathbf{tr}\left(\hat{Y}^T (\tilde{A}_k - A)^T \hat{Y}\hat{Y}^T (\tilde{A}_l - A)\hat{Y}\right)\right]
\end{aligned}$$

Since $\mathbf{E}\left[\tilde{A}\right] = A$, and the $\tilde{A}_l$ are independently sampled, the summand here will be zero unless $k = l$. Therefore,

$$\begin{aligned}
&\mathbf{E}\left[\left\|\bar{R} - \hat{Y}^T A \hat{Y}\right\|_F^2\right] \\
&= \frac{1}{L^2}\sum_{l=0}^{L-1} \mathbf{E}\left[\mathbf{tr}\left(\hat{Y}^T (\tilde{A}_l - A)^T \hat{Y}\hat{Y}^T (\tilde{A}_l - A)\hat{Y}\right)\right] \\
&= \frac{1}{L}\mathbf{E}\left[\mathbf{tr}\left(\hat{Y}^T (\tilde{A} - A)^T \hat{Y}\hat{Y}^T (\tilde{A} - A)\hat{Y}\right)\right] \\
&\leq \frac{1}{L}\mathbf{E}\left[\mathbf{tr}\left(\hat{Y}^T \tilde{A}^T \hat{Y}\hat{Y}^T \tilde{A}\hat{Y}\right)\right].
\end{aligned}$$

Applying the Alecton variance condition, and recalling that $\mathbf{tr}\left(\hat{Y}\hat{Y}^T\right) = p$, results in

$$\mathbf{E}\left[\left\|\bar{R} - \hat{Y}^T A \hat{Y}\right\|_F^2\right] \leq \frac{p^2\sigma_r^2}{L}.$$

We can now apply Markov's inequality to this expression. This results in, for any constant $\psi > 0$,

$$P\left(\left\|\bar{R} - \hat{Y}^T A \hat{Y}\right\|_F^2 \geq \psi\right) \leq \frac{p^2\sigma_r^2}{L\psi},$$

which is the desired result. □

D Proofs of Lemmas

First, we prove the lemmas used above to demonstrate the general result.

Proof of quadratic rational lower bound lemma (Lemma 6). Expanding the product results in

$$\begin{aligned}
\left(1+bx+cx^2\right)\left(1+(2a-b)x-cx^2\right) &= 1+((2a-b)+b)x+(c-c+(2a-b)b)x^2+((2a-b)c-bc)x^3-c^2x^4 \\
&= 1 + 2ax + (2ab - b^2)x^2 + 2(a-b)cx^3 - c^2x^4 \\
&= 1 + 2ax + a^2x^2 - (a^2 - 2ab + b^2)x^2 + 2(a-b)cx^3 - c^2x^4 \\
&= 1 + 2ax + a^2x^2 - x^2\left((a-b)^2 - 2(a-b)cx + c^2x^2\right) \\
&= (1+ax)^2 - x^2((a-b) - cx)^2 \\
&\leq (1+ax)^2.
\end{aligned}$$

Dividing both sides by $1 + bx + cx^2$ (which we can do since this is assumed to be positive) reconstructs the desired identity. □

Proof of Lemma 7. We first note that, by the symmetry of the multivariate Gaussian distribution, initializing Y_0 uniformly at random such that $Y_0^T Y_0 = I$ is equivalent to initializing the entries of Y_0 as independent standard normal random variables, for the purposes of computing τ_0. Under this initialization strategy, $\mathbf{E}[\tau_0]$ is

$$\begin{aligned}
\mathbf{E}[\tau_0] &= \mathbf{E}\left[\frac{|Y_0^T U Y_0|}{|Y_0^T W Y_0|}\right] \\
&= \mathbf{E}\left[\frac{|Y_0^T U Y_0|}{|\gamma n^{-1} p^{-2} q Y_0^T (I-U) Y_0 + Y_0^T U Y_0|}\right].
\end{aligned}$$

Now, let $X \in \mathbb{R}^{q\times p}$ be the component of Y_0 that is in the column space of U, and let $Z \in \mathbb{R}^{(n-q)\times p}$ be the component of Y_0 in the null space of U. Then,

$$\mathbf{E}[\tau_0] = \mathbf{E}\left[\frac{|X^T X|}{|\gamma n^{-1} p^{-2} q Z^T Z + X^T X|}\right].$$

Since X and Z are selected orthogonally from a Gaussian random matrix, they must be independent, so we can take their expected values independently. Taking the expected value first with respect to Z, we notice

that $|V|^{-1}$ is a convex function in V, and so by Jensen's inequality,

$$\begin{aligned}\mathbf{E}[\tau_0] &\geq \mathbf{E}\left[\frac{\left|X^TX\right|}{|\gamma n^{-1}p^{-2}q\mathbf{E}[Z^TZ]+X^TX|}\right]\\ &\geq \mathbf{E}\left[\frac{\left|X^TX\right|}{|\gamma n^{-1}p^{-2}q(n-q)I+X^TX|}\right]\\ &\geq \mathbf{E}\left[\frac{\left|X^TX\right|}{|\gamma p^{-2}qI+X^TX|}\right]\\ &= \mathbf{E}\left[\left|I+\gamma p^{-2}q(X^TX)^{-1}\right|^{-1}\right].\end{aligned}$$

Now, let $V \in \mathbb{R}^{q\times p}$ be a random full-rank projection matrix, selected independently of X. Then,

$$\mathbf{E}\left[VV^T\right] = \frac{p}{q}I,$$

and so

$$\mathbf{E}[\tau_0] \geq \mathbf{E}\left[\left|I+\gamma p^{-1}\mathbf{E}\left[X^TVV^TX\middle|X\right]^{-1}\right|^{-1}\right].$$

Applying Jensen's inequality again,

$$\mathbf{E}[\tau_0] \geq \mathbf{E}\left[\mathbf{E}\left[\left|I+\gamma p^{-1}\left(X^TVV^TX\right)^{-1}\right|^{-1}\middle|X\right]\right].$$

and by the law of total expectation,

$$\mathbf{E}[\tau_0] \geq \mathbf{E}\left[\left|I+\gamma p^{-1}\left(X^TVV^TX\right)^{-1}\right|^{-1}\right].$$

Now, since V and X were sampled independently, it follows that V^TX is sampled as a standard normal random matrix in $\mathbb{R}^{p\times p}$. If we call this matrix R, then

$$\begin{aligned}\mathbf{E}[\tau_0] &\geq \mathbf{E}\left[\left|I+\gamma p^{-1}\left(R^TR\right)^{-1}\right|^{-1}\right]\\ &= 1-\frac{1}{2}Z_p(\gamma),\end{aligned}$$

as desired. □

Lemma 15. *For any $B \in \mathbb{R}^{n\times n}$, any $Y \in \mathbb{R}^{n\times m}$, and any symmetric positive- semidefinite $Z \in \mathbb{R}^{n\times n}$, if either B is rank-1 or $m = 1$, then*

$$\begin{aligned}&\left|Y^T(I+B)^TZ(I+B)Y\right|\\ &\quad= \left|Y^TZY\right|\Big(\left(\mathbf{tr}\left(Y(Y^TZY)^{-1}Y^TZB\right)+1\right)^2\\ &\qquad+\mathbf{tr}\left(Y(Y^TZY)^{-1}Y^TB^TZB\right)\\ &\qquad-\mathbf{tr}\left(ZY(Y^TZY)^{-1}Y^TZBY(Y^TZY)^{-1}Y^TB^T\right)\Big).\end{aligned}$$

Proof. We will prove this separately for each case. First, if $m = 1$, then Y is a vector, and the desired expression simplifies to

$$\begin{aligned}
&Y^T(I+B)^T Z(I+B)Y \\
&\quad = Y^T ZY \left((Y^T ZY)^{-1} Y^T ZBY + 1\right)^2 \\
&\qquad + \mathbf{tr}\left(Y^T B^T ZBY\right) \\
&\qquad - (Y^T ZY)^{-1}(Y^T ZBY)^2.
\end{aligned}$$

Straightforward evaluation indicates that this expression holds in this case.

Next, we consider the case where B is rank-1. In this case, we can rewrite it as $B = uv^T$ for vectors u and v, such that $u^T Zu = 1$. Then,

$$\begin{aligned}
&\left|Y^T(I+B)^T Z(I+B)Y\right| \\
&\quad = \left|Y^T(I+uv^T)^T Z(I+uv^T)Y\right| \\
&\quad = \left|Y^T ZY + 2Y^T Zuv^T Y + Y^T vv^T Y\right|
\end{aligned}$$

If we define $M = Y^T ZY$ and

$$W = \begin{bmatrix} Y^T Zu & Y^T v \end{bmatrix},$$

then

$$\begin{aligned}
&\left|Y^T(I+B)^T Z(I+B)Y\right| \\
&\quad = \left|M + W \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix} W^T\right|.
\end{aligned}$$

Applying the matrix determinant lemma, and recalling that

$$\begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix}^{-1} = \begin{bmatrix} -1 & 1 \\ 1 & 0 \end{bmatrix}$$

and

$$\begin{vmatrix} 0 & 1 \\ 1 & 1 \end{vmatrix} = -1,$$

we produce

$$\begin{aligned}
&-\det M^{-1}\left|Y^T(I+B)^T Z(I+B)Y\right| \\
&\quad = -\left|\begin{bmatrix} -1 & 1 \\ 1 & 0 \end{bmatrix} + W^T M^{-1} W\right| \\
&\quad = \begin{vmatrix} u^T ZYM^{-1}Y^T Zu - 1 & v^T YM^{-1}Y^T Zu + 1 \\ v^T YM^{-1}Y^T Zu + 1 & v^T YM^{-1}Y^T v \end{vmatrix} \\
&\quad = \left(u^T ZYM^{-1}Y^T Zu - 1\right)\left(v^T YM^{-1}Y^T v\right) \\
&\qquad - \left(v^T YM^{-1}Y^T Zu + 1\right)^2 \\
&\quad = u^T ZYM^{-1}Y^T Zuv^T YM^{-1}Y^T v \\
&\qquad - v^T YM^{-1}Y^T vu^T Zu \\
&\qquad - \left(v^T YM^{-1}Y^T Zu + 1\right)^2.
\end{aligned}$$

Rewriting this in terms of the matrix $B = uv^T$,

$$\begin{aligned}
&- \det M^{-1} \left| Y^T (I+B)^T Z (I+B) Y \right| \\
&\quad = \mathbf{tr}\left(ZYM^{-1}Y^T ZBYM^{-1}Y^T B^T \right) \\
&\qquad - \mathbf{tr}\left(YM^{-1}Y^T B^T ZB \right) \\
&\qquad - \left(\mathbf{tr}\left(YM^{-1}Y^T ZB \right) + 1 \right)^2 .
\end{aligned}$$

Substitution produces the desired result. □

Proof of Lemma 8. First, for the lower bound, we notice that

$$ZY(Y^T ZY)^{-1}Y^T Z \preceq Z,$$

since the interior of the left expression is a projection matrix. This lets us conclude that

$$\begin{aligned}
&\mathbf{tr}\left(Y(Y^T ZY)^{-1}Y^T B^T ZB \right) \\
&\quad \geq \mathbf{tr}\left(ZY(Y^T ZY)^{-1}Y^T ZBY(Y^T ZY)^{-1}Y^T B^T \right).
\end{aligned}$$

Appling this to the result of Lemma 15 produces the desired lower bound.

For the upper bound, recall that, by the Cauchy-Schwarz inequality, for any rank-1 matrix A,

$$\mathbf{tr}\left(A\right)^2 \leq \mathbf{tr}\left(A^T A\right).$$

Since B is rank-1, it follows that

$$\begin{aligned}
&\mathbf{tr}\left(Y(Y^T ZY)^{-1}Y^T ZB \right) \\
&\quad \leq \mathbf{tr}\left(ZY(Y^T ZY)^{-1}Y^T ZBY(Y^T ZY)^{-1}Y^T B^T \right).
\end{aligned}$$

Appling this to the result of Lemma 15 produces the desired upper bound. □

Lemma 16. *For any symmetric matrix* $0 \preceq X \preceq I$*,*

$$\mathbf{tr}\left(I - X\right) \geq 1 - \left|X\right|.$$

Proof. If $x_1, x_2, \ldots, x_p$ are the eigenvalues of x, then this statement is equivalent to

$$\left(\sum_{i=1}^{p} (1 - x_i) \right) - \left(1 - \prod_{i=1}^{p} x_i \right) > 0.$$

If we let $f(X)$ denote this expression, then

$$\frac{\partial f}{\partial x_j} = -1 + \frac{1}{x_j} \prod_{i=1}^{p} x_i \leq 0.$$

It follows that the minimum of f is attained at $X = I$. However, when $X = I$, $f(X) = 0$, and so $f > 0$, which proves the lemma. □

Proof of Lemma 10. From the definition of ϕ_k, if we let $Z^2 = \left(Y_k^T W Y_k\right)^{-1}$ for Z positive semidefinite, then

$$\begin{aligned}\phi_k &= \mathbf{tr}\left(I - Y_k^T U^T U Y_k \left(Y_k^T W Y_k\right)^{-1}\right) \\ &= \mathbf{tr}\left(I - Z Y_k^T U^T U Y_k Z\right).\end{aligned}$$

Since $0 \preceq Z Y_k^T U^T U Y_k Z \preceq I$, we can apply Lemma 16, which produces

$$\begin{aligned}\phi_k &\geq 1 - \left|Z Y_k^T U^T U Y_k Z\right| \\ &= 1 - \frac{\left|Y_k^T U^T U Y_k\right|}{\left|Y_k^T W Y_k\right|} \\ &= 1 - \tau_k,\end{aligned}$$

which is the desired expression. □

Proof of Lemma 9. Since the success event does not occur, it follows that there exists a $z \in \mathbb{R}^p$ such that

$$\frac{\|U Y_k z\|^2}{\|Y_k z\|^2} \leq 1 - \epsilon.$$

If we let

$$\hat{Y}_k = Y_k \left(Y_k^T Y_k\right)^{-\frac{1}{2}},$$

and define $\hat{z}$ as the unit vector such that

$$\hat{z} \propto \left(Y_k^T Y_k\right)^{\frac{1}{2}} z,$$

then we can rewrite this as

$$\left\|U \hat{Y}_k \hat{z}\right\|^2 \leq 1 - \epsilon.$$

It follows that $\hat{Y}_k^T U \hat{Y}_k$ has an eigenvalues less than $1 - \epsilon$.

Now, expanding τ_k,

$$\begin{aligned}\tau_k &= \frac{\left|Y_k^T U Y_k\right|}{\left|Y_k^T W Y_k\right|} \\ &= \frac{\left|\hat{Y}_k^T U \hat{Y}_k\right|}{\left|\hat{Y}_k^T W \hat{Y}_k\right|} \\ &= \left|(1 - \gamma n^{-1} p^{-2} q) I + \gamma n^{-1} p^{-2} q \left(\hat{Y}_k^T U \hat{Y}_k\right)^{-1}\right|^{-1}\end{aligned}$$

Since this is a matrix that has eigenvalues between 0 and 1, it follows that its determinant is less than each of its eigenvalues. From the analysis above, we can bound one of the eigenvalues of this matrix. Doing this results in

$$\begin{aligned}\tau_k &\leq \left((1 - \gamma n^{-1} p^{-2} q) + \gamma n^{-1} p^{-2} q \left(1 - \epsilon\right)^{-1}\right)^{-1} \\ &= \frac{1 - \epsilon}{\gamma n^{-1} p^{-2} q + (1 - \gamma n^{-1} p^{-2} q)(1 - \epsilon)} \\ &= 1 - \frac{\gamma n^{-1} p^{-2} q \epsilon}{\gamma n^{-1} p^{-2} q + (1 - \gamma n^{-1} p^{-2} q)(1 - \epsilon)} \\ &\leq 1 - \gamma n^{-1} p^{-2} q \epsilon,\end{aligned}$$

as desired. □

Lemma 17. *Let x be a standard normal random variable, and $a \in \mathbb{R}$ a constant. Then*

$$\mathbf{E}\left[\frac{a^2}{x^2+a^2}\right] = \exp\left(\frac{a^2}{2}\right)\sqrt{\frac{\pi a^2}{2}}\operatorname{erfc}\left(\sqrt{\frac{a^2}{2}}\right).$$

Proof. By the definition of expected value, since x is normally distributed,

$$\mathbf{E}\left[\frac{a^2}{x^2+a^2}\right] = \int_{-\infty}^{\infty}\left(\frac{a^2}{x^2+a^2}\right)\left(\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right)\right)dx.$$

If we let $\mathcal{F}$ denote the fourier transform, then

$$\mathcal{F}\left[\frac{a}{x^2+a^2}\right] = \sqrt{2\pi}\exp\left(-a\left|\omega\right|\right).$$

Furthermore, since the Gaussian functions are eigenfunctions of the Fourier transform, we know that

$$\mathcal{F}\left[\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right)\right] = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{\omega^2}{2}\right).$$

And so, by Parseval's theorem,

$$\begin{aligned}
\mathbf{E}\left[\frac{1}{x^2+1}\right] &= a\int_{-\infty}^{\infty}\mathcal{F}\left[\frac{a}{x^2+a^2}\right]\mathcal{F}\left[\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right)\right]d\omega \\
&= a\int_{-\infty}^{\infty}\sqrt{2\pi}\exp\left(-a\left|\omega\right|\right)\left(\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{\omega^2}{2}\right)\right)d\omega \\
&= a\int_{0}^{\infty}\exp\left(-a\omega-\frac{\omega^2}{2}\right)d\omega \\
&= a\exp\left(\frac{a^2}{2}\right)\int_{0}^{\infty}\exp\left(-\frac{a^2}{2}-a\omega-\frac{\omega^2}{2}\right)d\omega.
\end{aligned}$$

Letting $u = \frac{\omega+a}{\sqrt{2}}$ and $d\omega = \sqrt{2}du$, so

$$\begin{aligned}
\mathbf{E}\left[\frac{1}{x^2+1}\right] &= a\exp\left(\frac{a^2}{2}\right)\int_{\frac{a}{\sqrt{2}}}^{\infty}\exp\left(-u^2\right)\sqrt{2}du \\
&= \exp\left(\frac{a^2}{2}\right)\sqrt{\frac{\pi a^2}{2}}\operatorname{erfc}\left(\sqrt{\frac{a^2}{2}}\right),
\end{aligned}$$

as desired. □

Proof of Lemma 1. We start by stating the definition of $Z_1(\gamma)$. For some Gaussian random matrix $R \in \mathbb{R}^{1\times 1}$,

$$Z_1(\gamma) = 2\left(1 - \mathbf{E}\left[\left|I+\gamma(R^TR)^{-1}\right|^{-1}\right]\right).$$

Since R is a scalar, this reduces to

$$\begin{aligned}
Z_1(\gamma) &= 2\left(1-\mathbf{E}\left[\left(1+\gamma R^{-2}\right)^{-1}\right]\right) \\
&= \mathbf{E}\left[2\left(1-\frac{1}{1+\gamma R^{-2}}\right)\right] \\
&= \mathbf{E}\left[2\frac{\gamma R^{-2}}{1+\gamma R^{-2}}\right] \\
&= 2\mathbf{E}\left[\left(\frac{\gamma}{R^2+\gamma}\right)\right].
\end{aligned}$$

Applying Lemma 17,

$$\begin{aligned} Z_1(\gamma) &= 2\exp\left(\frac{\gamma}{2}\right)\sqrt{\frac{\pi\gamma}{2}}\operatorname{erfc}\left(\sqrt{\frac{\gamma}{2}}\right) \\ &= \sqrt{2\pi\gamma}\exp\left(\frac{\gamma}{2}\right)\operatorname{erfc}\left(\sqrt{\frac{\gamma}{2}}\right). \end{aligned}$$

This is the desired expression. Furthermore, since for all x,

$$\operatorname{erfc}\left(\sqrt{x}\right) \leq \exp\left(x\right),$$

we can also produce the desired upper bound on Z_1,

$$Z_1 \leq \sqrt{2\pi\gamma}.$$

□

D.1 Proofs of Alecton Variance Condition Lemmas

Next, we prove the Alecton Variance Conditions lemmas for the distributions mentioned in the body of the paper.

D.1.1 Entrywise Sampling

To analyze the entrywise sampling case, we need some lemmas that makes the incoherence condition more accessible.

Lemma 18. *If matrix A is symmetric and incoherent with parameter μ, and B is a symmetric matrix that commutes with A, then B is incoherent with parameter μ.*

Proof. Since A and B commute, they must have the same eigenvectors. Therefore, the set of eigenvectors that shows that A is incoherent with parameter μ will also show that B has the same property. □

Lemma 19. *If matrix A is symmetric and incoherent with parameter μ, and e_i is a standard basis element, then*

$$e_i^T A e_i \leq \frac{\mu^2}{n}\mathbf{tr}\left(A\right).$$

Proof. Let $u_1, u_2, \ldots, u_n$ be the eigenvectors guaranteed by the incoherence of A, and let $\lambda_1, \ldots, \lambda_n$ be the corresponding eigenvalues. Then,

$$\begin{aligned} e_i^T A e_i &= e_i^T \left(\sum_{j=1}^{n} u_j \lambda_j u_j^T\right) e_i \\ &= \sum_{j=1}^{n} u_j \lambda_j (e_i^T u_j)^2. \end{aligned}$$

Applying the definition of incoherence,

$$e_i^T A e_i \leq \sum_{j=1}^{n} u_j \lambda_j \left(\frac{\mu}{\sqrt{n}}\right)^2 = \frac{\mu^2}{n}\mathbf{tr}\left(A\right),$$

as desired. □

Proof of the σ_a bound part of Lemma 2. We recall that the entrywise samples are of the form

$$\tilde{A} = n^2 uu^T Avv^T,$$

where u and v are independently, uniformly chosen standard basis elements. We further recall that $\mathbf{E}\left[uu^T\right] = \mathbf{E}\left[vv^T\right] = n^{-1}I$. Now, evaluating the desired quantity,

$$\mathbf{E}\left[y^T\tilde{A}^T W\tilde{A}y\right] = n^4\mathbf{E}\left[y^T vv^T Auu^T Wuu^T Avv^T y\right].$$

Since W commutes with A, by Lemmas 18 and 19, $u^T Wu \leq \mu^2 n^{-1}\mathbf{tr}\left(W\right)$. Therefore,

$$\begin{aligned}\mathbf{E}\left[y^T\tilde{A}^T W\tilde{A}y\right] &\leq \mu^2 n^3\mathbf{tr}\left(W\right)\mathbf{E}\left[y^T vv^T Auu^T Avv^T y\right]\\ &= \mu^2 n^2\mathbf{tr}\left(W\right)\mathbf{E}\left[y^T vv^T A^2 vv^T y\right].\end{aligned}$$

Since A^2 commutes with A, the same logic shows that $v^T A^2 v \leq \mu^2 n^{-1}\mathbf{tr}\left(A^2\right)$, and so,

$$\begin{aligned}\mathbf{E}\left[y^T\tilde{A}^T W\tilde{A}y\right] &\leq \mu^4 n\mathbf{tr}\left(W\right)\mathbf{tr}\left(A^2\right)\mathbf{E}\left[y^T vv^T y\right]\\ &= \mu^4\mathbf{tr}\left(W\right)\left\|A\right\|_F^2\left\|y\right\|^2.\end{aligned}$$

So it suffices to choose $\sigma_a^2 = \mu^4\left\|A\right\|_F^2$, as desired. □

Proof of the σ_r bound part of Lemma 2. Evaluating the desired quantity,

$$\begin{aligned}\mathbf{E}\left[\left(y^T\tilde{A}y\right)^2\right] &= n^4\mathbf{E}\left[\left(y^T uu^T Avv^T y\right)^2\right]\\ &= n^4\mathbf{E}\left[(u^T y)^2(v^T y)^2(u^T Av)^2\right].\end{aligned}$$

By the CauchySchwarz inequality,

$$(u^T Av)^2 \leq (u^T Au)(v^T Av),$$

and by Lemma 19, $u^T Au \leq \mu^2 n^{-1}\mathbf{tr}\left(A\right)$, and so

$$(u^T Av)^2 \leq \mu^4 n^{-2}\mathbf{tr}\left(A\right)^2.$$

Therefore,

$$\begin{aligned}\mathbf{E}\left[\left(y^T\tilde{A}y\right)^2\right] &\leq \mu^4 n^2\mathbf{tr}\left(A\right)^2\mathbf{E}\left[(u^T y)^2(v^T y)^2\right]\\ &= \mu^4\mathbf{tr}\left(A\right)^2\left\|y\right\|^4.\end{aligned}$$

So it suffices to choose $\sigma_r^2 = \mu^4\mathbf{tr}\left(A\right)^2$, as desired. □

D.1.2 Rectangular Entrywise Sampling

Proof of Lemma 3. We recall that the rectangular entrywise samples are of the form

$$\tilde{A} = mnM_{ij}(e_i e_{m+j}^T + e_{m+j}e_i^T),$$

where $i \in 1, \ldots, m$ and $j \in 1, \ldots, n$ are chosen uniformly and independently. Now, for any y and z in $\mathbb{R}^{m+n}$,

$$\mathbf{E}\left[(z^T \tilde{A} y)^2\right] = m^2 n^2 \mathbf{E}\left[M_{ij}^2 (z^T (e_i e_{m+j}^T + e_{m+j} e_i^T) y)^2\right].$$

Applying the entry bound,

$$\mathbf{E}\left[(z^T \tilde{A} y)^2\right] \leq \xi mn \left\|M\right\|_F^2 \mathbf{E}\left[(z^T e_i e_{m+j}^T y + z^T e_{m+j} e_i^T y)^2\right].$$

Now, since $(x+y)^2 \leq 2(x^2+y^2)$, if we let P be the projection matrix onto the first m basis vectors, then $\mathbf{E}\left[e_i e_i^T\right] = m^{-1} P$ and $\mathbf{E}\left[e_{m+j} e_{m+j}^T\right] = n^{-1}(I-P)$, and so,

$$\begin{aligned}
&\mathbf{E}\left[(z^T \tilde{A} y)^2\right] \\
&\leq 2\xi mn \left\|M\right\|_F^2 \mathbf{E}\left[(z^T e_i)^2 (e_{m+j}^T y)^2 + (z^T e_{m+j})^2 (e_i^T y)^2\right] \\
&= 2\xi \left\|M\right\|_F^2 \left(\left\|Pz\right\|^2 \left\|(I-P)y\right\|^2 + \left\|(I-P)z\right\|^2 \left\|Py\right\|^2\right) \\
&\leq 2\xi \left\|M\right\|_F^2 \left\|y\right\|^2 \left\|z\right\|^2.
\end{aligned}$$

Since this is true for any y and z, it is true in particular for z being an eigenvector of A. Therefore, it suffices to pick $\sigma_a^2 = 2\xi \left\|M\right\|_F^2$. Similarly, it is true in particular for $z = y$, and therefore it suffices to pick $\sigma_r^2 = 2\xi \left\|M\right\|_F^2$. This proves the lemma. □

D.1.3 Trace Sampling

In order to prove our second moment lemma for the trace sampling case, we must first derive some lemmas about the way this distribution behaves.

Lemma 20 (Sphere Component Fourth Moment)**.** *If $n > 50$, and $v \in \mathbb{R}^n$ is sampled uniformly from the unit sphere, then for any unit vector $y \in \mathbb{R}^n$,*

$$\mathbf{E}\left[\left(y^T v\right)^4\right] \leq \frac{4}{n^2}.$$

Proof. Let x be sampled from the standard normal distribution in $\mathbb{R}^n$. Then, by radial symmetry,

$$\mathbf{E}\left[\left(y^T v\right)^4\right] = \mathbf{E}\left[\frac{\left(y^T x\right)^4}{\left\|x\right\|^4}\right].$$

If we let u denote $y^T x$, and z denote the components of x orthogonal to y, then $\left\|x\right\|^2 = u^2 + \left\|z\right\|^2$. Furthermore, by the properties of the normal distribution, u and z are independent. Therefore,

$$\begin{aligned}
\mathbf{E}\left[\left(y^T v\right)^4\right] &= \mathbf{E}\left[u^4 \left(u^2 + \left\|z\right\|^2\right)^{-2}\right] \\
&\leq \mathbf{E}\left[u^4 \left(\left\|z\right\|^2\right)^{-2}\right] \\
&= \mathbf{E}\left[u^4\right] \mathbf{E}\left[\left\|z\right\|^{-4}\right].
\end{aligned}$$

Now, $\mathbf{E}\left[u^4\right]$ is the fourth moment of the normal distribution, which is known to be 3. Furthermore, $\mathbf{E}\left[\|z\|^{-4}\right]$ is the second moment of an inverse-chi-squared distribution with parameter $n-1$, which is also a known result. Substituting these in,

$$\begin{aligned}\mathbf{E}\left[\left(y^T v\right)^4\right] &\leq 3\left((n-3)^{-2} + 2\left(n-3\right)^{-2}\left(n-5\right)^{-1}\right) \\ &= 3\left(n-3\right)^{-2}\left(1 + 2\left(n-5\right)^{-1}\right).\end{aligned}$$

This quantity has the asymptotic properties we want. In particular, applying the constraint that $n > 50$,

$$\mathbf{E}\left[\left(y^T v\right)^4\right] \leq \frac{4}{n^2}.$$

This is the desired result. □

Lemma 21 (Sphere Component Fourth Moment Matrix)**.** *If $n > 50$, and $v \in \mathbb{R}^n$ is sampled uniformly from the unit sphere, then for any positive semidefinite matrix W,*

$$\mathbf{E}\left[vv^T W vv^T\right] \preceq 4n^{-2}\mathbf{tr}\left(W\right) I.$$

Proof. Let

$$W = \sum_{i=1}^{n} \lambda_i w_i w_i^T$$

be the eigendecomposition of W. Then for any unit vector z,

$$\begin{aligned}z^T \mathbf{E}\left[vv^T W vv^T\right] z &= \mathbf{E}\left[z^T vv^T \left(\sum_{i=1}^{n} \lambda_i w_i w_i^T\right) vv^T z\right] \\ &= \sum_{i=1}^{n} \lambda_i \mathbf{E}\left[\left(z^T v\right)^2 \left(w_i^T v\right)^2\right].\end{aligned}$$

By the Cauchy-Schwarz inequality applied to the expectation,

$$\begin{aligned}\mathbf{E}\left[\left(z^T v\right)^2 \left(w_i^T v\right)^2\right] &\leq \sqrt{\mathbf{E}\left[\left(z^T v\right)^4\right] \mathbf{E}\left[\left(w_i^T v\right)^2\right]} \\ &= \mathbf{E}\left[\left(z^T v\right)^4\right].\end{aligned}$$

By Lemma 20, $\mathbf{E}\left[\left(z^T v\right)^4\right] \leq 4n^{-2}$, and so

$$z^T \mathbf{E}\left[vv^T W vv^T\right] z \leq \sum_{i=1}^{n} \lambda_i (4n^{-2}) = 4n^{-2}\mathbf{tr}\left(W\right).$$

Since this is true for any unit vector z, by the definition of the positive semidefinite relation,

$$\mathbf{E}\left[vv^T W vv^T\right] \preceq 4n^{-2}\mathbf{tr}\left(W\right) I,$$

as desired. □

Now, we prove the AVC lemma for this distribution.

Proof of σ_a bound part of Lemma 4. Evaluating the expression we want to bound,

$$\mathbf{E}\left[y^T\tilde{A}^T W\tilde{A}y\right] = n^4\mathbf{E}\left[y^T vv^T Auu^T Wuu^T Avv^T y\right].$$

Applying Lemma 21,

$$\begin{aligned}\mathbf{E}\left[y^T\tilde{A}^T W\tilde{A}y\right] &\leq n^4\mathbf{E}\left[y^T vv^T A\left(4n^{-2}\mathbf{tr}\left(W\right)I\right)Avv^T y\right]\\ &= 4n^2\mathbf{tr}\left(W\right)\mathbf{E}\left[y^T vv^T A^2 vv^T y\right].\end{aligned}$$

Again applying Lemma 21,

$$\begin{aligned}\mathbf{E}\left[y^T\tilde{A}^T W\tilde{A}y\right] &\leq 4n^2\mathbf{tr}\left(W\right)y^T\left(4n^{-2}\mathbf{tr}\left(A^2\right)I\right)y\\ &= 16\left\|A\right\|_F^2\mathbf{tr}\left(W\right)\left\|y\right\|^2.\end{aligned}$$

So it suffices to pick $\sigma_a^2 = 16\left\|A\right\|_F^2$, as desired. □

Proof of σ_r bound part of Lemma 4. Evaluating the expression we want to bound,

$$\begin{aligned}\mathbf{E}\left[\left(y\tilde{A}y\right)^2\right] &= n^4\mathbf{E}\left[\left(yvv^T Aww^T y\right)^2\right]\\ &= n^4\mathbf{E}\left[\mathbf{tr}\left(Avv^T yy^T vv^T Aww^T yy^T ww^T\right)\right]\\ &= n^4\mathbf{tr}\left(A\mathbf{E}\left[vv^T yy^T vv^T\right]A\mathbf{E}\left[ww^T yy^T ww^T\right]\right).\end{aligned}$$

Applying Lemma 21 to this results in

$$\begin{aligned}&\mathbf{E}\left[\left(y\tilde{A}y\right)^2\right]\\ &\leq n^4\mathbf{tr}\left(A\left(4n^{-2}\mathbf{tr}\left(yy^T\right)I\right)A\left(4n^{-2}\mathbf{tr}\left(yy^T\right)I\right)\right)\\ &= 16\left\|A\right\|_F^2\left\|y\right\|^4.\end{aligned}$$

So it suffices to pick $\sigma_r^2 = 16\left\|A\right\|_F^2$, as desired. □

D.1.4 Subspace Sampling

Recall that, in subspace sampling, our samples are of the form

$$\tilde{A} = rn^2m^{-2}Qvv^T R,$$

where Q and R are independent projection matrices that select m entries uniformly at random, and v is uniformly and independently selected from the column space of A. Using this, we first prove some lemmas, then prove our bounds.

Lemma 22. *If Q is a projection matrix that projects onto a subspace spanned by m random standard basis vectors, and v is a member of a subspace that is incoherent with parameter μ, then for any vector x,*

$$(x^T Qv)^2 \leq (\mu mr + m^2)n^{-2}\left\|x\right\|^2\left\|v\right\|^2.$$

As a corollary, for any symmetric matrix $W \succeq 0$,

$$v^T QWQv \leq (\mu mr + m^2)n^{-2}\mathbf{tr}\left(W\right)\left\|v\right\|^2.$$

Proof. Let λ_i be 1 in the event that e_i is in the column space of Q, and 0 otherwise. Then an eigendecomposition of Q is

$$Q = \sum_{i=1}^{n} \lambda_i e_i e_i^T.$$

Therefore,

$$\begin{aligned}(x^T Q v)^2 &= \left(\sum_{i=1}^{n} \lambda_i x^T e_i e_i^T v\right)^2 \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j x_i x_j v_i v_j.\end{aligned}$$

Taking the expected value, and noting that λ_i and λ_j are independent, and have expected value $\mathbf{E}[\lambda_i] = mn^{-1}$,

$$\begin{aligned}\mathbf{E}\left[(x^T Q v)^2\right] = m^2 n^{-2} &\sum_{i=1}^{n}\sum_{j=1}^{n} x_i x_j v_i v_j \\ &+ mn^{-1}(1 - mn^{-1}) \sum_{i=1}^{n} x_i^2 v_i^2.\end{aligned}$$

Since v is part of a subspace that is incoherent,

$$\begin{aligned}\mathbf{E}\left[(x^T Q v)^2\right] &\leq m^2 n^{-2} \sum_{i=1}^{n}\sum_{j=1}^{n} x_i x_j v_i v_j \\ &\quad + \mu m r n^{-2}(1 - mn^{-1}) \|v\|^2 \sum_{i=1}^{n} x_i^2 \\ &= m^2 n^{-2} (x^T v)^2 \\ &\quad + \mu m r n^{-2} \|x\|^2 \|v\|^2 \\ &\leq (\mu m r + m^2) n^{-2} \|x\|^2 \|v\|^2,\end{aligned}$$

as desired. □

Proof of σ_a bound part of Lemma 5. Evaluating the expression we want to bound,

$$\begin{aligned}&\mathbf{E}\left[y^T \tilde{A}^T W \tilde{A} y\right] \\ &\quad = r^2 n^4 m^{-4} \mathbf{E}\left[y^T R v v^T Q W Q v v^T R y\right] \\ &\quad = r^2 n^4 m^{-4} \mathbf{E}\left[\mathbf{E}\left[v^T R y y^T R v\right] \mathbf{E}\left[v^T Q W Q v\right]\right].\end{aligned}$$

Applying Lemma 22,

$$\begin{aligned}\mathbf{E}\left[y^T \tilde{A}^T W \tilde{A} y\right] &\leq r^2 m^{-4} (\mu m r + m^2)^2 \mathbf{tr}(W) \|y\|^2 \\ &= r^2 (1 + \mu r m^{-1})^2 \mathbf{tr}(W) \|y\|^2.\end{aligned}$$

So, we can choose $\sigma_a^2 = r^2(1 + \mu r m^{-1})^2$, as desired. □

Proof of σ_r bound part of Lemma 5. Evaluating the expression we want to bound,

$$\begin{aligned}\mathbf{E}\left[(y^T\tilde{A}y)^2\right] &= r^2n^4m^{-4}\mathbf{E}\left[(y^TQvv^TRy)^2\right] \\ &= r^2n^4m^{-4}\mathbf{E}\left[\mathbf{E}\left[(y^TQv)^2\right]\mathbf{E}\left[(y^TRv)^2\right]\right].\end{aligned}$$

Applying Lemma 22,

$$\begin{aligned}\mathbf{E}\left[(y^T\tilde{A}y)^2\right] &\leq r^2m^{-4}(\mu mr+m^2)^2\left\|y\right\|^4 \\ &= r^2(1+\mu rm^{-1})^2\left\|y\right\|^4.\end{aligned}$$

So, we can choose $\sigma_r^2 = r^2(1+\mu rm^{-1})^2$, as desired. □

E Lower Bound on Alecton Rate

In this section, we prove a rough lower bound on the rate of convergence of an Alecton-like algorithm for bounded sampling distributions. Specifically, we analyze the case where, rather than choosing a constant η, we allow the step size to vary at each timestep. Our result shows that we can't hope for a better step size rule that improves the convergence rate of Alecton to, for example, a linear rate.

To show this lower bound, we assume we run Alecton with $p = 1$ for some sampling distribution such that for all η and all y, for some constant C,

$$\left\|y+\eta\tilde{A}y\right\| \leq (1+\eta C)\left\|y\right\|.$$

Further assume that for some eigenvector u (with eigenvalue $\lambda \geq 0$) that is not global solution, the sample variance in the direction of u satisfies

$$\mathbf{E}\left[\tilde{A}^Tuu^T\tilde{A}\right] \geq \sigma^2 I.$$

We now define ρ_k to be

$$\rho_k = \frac{(u^TY_k)^2}{\left\|Y_k\right\|^2}.$$

This quantity measures the error of the iterate at timestep k in the direction of u. We will show that the expected value of ρ_k can only decrease with at best a $\Omega\left(\frac{1}{K+1}\right)$ rate.

First, we require a lemma.

Lemma 23. *For any $a \geq 0$, $b \geq 0$, and $0 \leq x \leq 1$,*

$$a(1-x)^2+bx^2 \geq \frac{ab}{a+b}.$$

Proof. Expanding the left side,

$$\begin{aligned}a(1-x)^2+bx^2 &= a-2ax+(a+b)x^2 \\ &= a-\frac{a^2}{a+b}+\frac{a^2}{a+b}-2ax+(a+b)x^2 \\ &= \frac{ab}{a+b}+\frac{(a-(a+b)x)^2}{a+b} \\ &\geq \frac{ab}{a+b},\end{aligned}$$

as desired. □

Theorem 3. *Under the above conditions, regardless of how we choose the step size in the Alecton algorithm, even if we are able to choose a different step size each iteration, the expected error will still satisfy*

$$\mathbf{E}\left[\rho_K\right] \geq \frac{\sigma^2}{\sigma^2 n + C^2 K}.$$

Proof. Using the Alecton update rule with a time-varying step size η_k,

$$\begin{aligned}
\rho_{k+1} &= \frac{(u^T Y_k)^2}{\|Y_k\|^2} \\
&= \frac{(u^T Y_k + \eta_k u^T \tilde{A}_k Y_k)^2}{\left\|Y_k + \eta_k \tilde{A}_k Y_k\right\|^2} \\
&\geq \frac{(u^T Y_k + \eta_k u^T \tilde{A}_k Y_k)^2}{(1+\eta_k C)^2 \|Y_k\|^2}.
\end{aligned}$$

Taking the expected value,

$$\begin{aligned}
\mathbf{E}\left[\rho_{k+1}\right] &\geq \mathbf{E}\left[\frac{(u^T Y_k + \eta_k u^T \tilde{A}_k Y_k)^2}{(1+\eta_k C)^2 \|Y_k\|^2}\right] \\
&\geq \mathbf{E}\left[\frac{(1+2\eta_k\lambda)(u^T Y_k)^2 + \eta_k^2 \sigma^2 Y_k^T Y_k}{(1+\eta_k C)^2 \|Y_k\|^2}\right] \\
&= \frac{1+2\eta_k\lambda}{(1+\eta_k C)^2}\mathbf{E}\left[\rho_k\right] + \frac{\eta_k^2\sigma^2}{(1+\eta_k C)^2} \\
&\geq \frac{1}{(1+\eta_k C)^2}\mathbf{E}\left[\rho_k\right] + \frac{\eta_k^2\sigma^2}{(1+\eta_k C)^2}
\end{aligned}$$

Now, if we define ζ_k as

$$\zeta_k = \frac{\eta_k C}{1+\eta_k C},$$

then

$$\mathbf{E}\left[\rho_{k+1}\right] \geq (1-\zeta_k)^2 \mathbf{E}\left[\rho_k\right] + \zeta_k^2 \sigma^2 C^{-2}.$$

Applying Lemma 23,

$$\mathbf{E}\left[\rho_{k+1}\right] \geq \frac{\sigma^2 C^{-2}\mathbf{E}\left[\rho_k\right]}{\mathbf{E}\left[\rho_k\right] + \sigma^2 C^{-2}}.$$

Taking the inverse,

$$\frac{1}{\mathbf{E}\left[\rho_{k+1}\right]} \leq \frac{1}{\mathbf{E}\left[\rho_k\right]} + \frac{C^2}{\sigma^2}.$$

Therefore, summing across steps,

$$\frac{1}{\mathbf{E}\left[\rho_K\right]} \leq \frac{1}{\mathbf{E}\left[\rho_0\right]} + \frac{C^2 K}{\sigma^2}.$$

Since, by symmetry, $\mathbf{E}\left[\rho_0\right] = n^{-1}$, we have

$$\frac{1}{\mathbf{E}\left[\rho_K\right]} \leq n + \frac{C^2 K}{\sigma^2}.$$

and taking the inverse again produces

$$\mathbf{E}\left[\rho_K\right] \geq \frac{\sigma^2}{\sigma^2 n + C^2 K},$$

which is the desired expression. □

F Handling Constraints

Alecton can easily be adapted to solve the problem of finding a low-rank approximation to a matrix under a *spectahedral constraint*. That is, we want to solve the problem

$$\begin{array}{ll}\text{minimize} & \|A - X\|_F^2 \\ \text{subject to} & X \in \mathbb{R}^{N\times N}, \mathbf{tr}\,(X) = 1, \\ & \mathbf{rank}\,(X) \leq 1, X \succeq 0.\end{array}$$

This is equivalent to the decomposed problem

$$\begin{array}{ll}\text{minimize} & \|y\|^4 - 2y^T Ay + \|A\|_F^2 \\ \text{subject to} & y \in \mathbb{R}^N, \|y\|^2 = 1,\end{array}$$

which is itself equivalent to:

$$\begin{array}{ll}\text{minimize} & 1 - 2y^T Ay + \|A\|_F^2 \\ \text{subject to} & y \in \mathbb{R}^N, \|y\|^2 = 1.\end{array}$$

This will have a minimum when $y = u_1$. We can therefore solve the problem using only the angular phase of Alecton, which recovers the vector u_1. The same convergence analysis described above still applies.

For an example of a constrained problem that Alecton cannot handle, because it is NP-hard, see the elliptope-constrained MAXCUT embedding in Appendix A. This shows that constrained problems can't be solved efficiently by SGD algorithms in all cases.

G Towards a Linear Rate

In this section, we consider a special case of the matrix recovery problem: one in which the samples we are given would allow us to exactly recover A. That is, for some linear operator $\Omega : \mathbb{R}^{n\times n} \to \mathbb{R}^s$, we are given the value of $\Omega(A)$ as an input, and we know that the unique solution of the optimization problem

$$\begin{array}{ll}\text{minimize} & \|\Omega(X - A)\|^2 \\ \text{subject to} & X \in \mathbb{R}^{n\times n}, \mathbf{rank}\,(X) \leq p, X \succeq 0\end{array}$$

is $X = A$. Performing a rank-p quadratic substitution on this problem results in:

$$\begin{array}{ll}\text{minimize} & \left\|\Omega(YY^T - A)\right\|^2 \\ \text{subject to} & Y \in \mathbb{R}^{n\times p}\end{array}$$

The specific case we will be looking at is where the operator Ω satisfies the p-RIP constraint.

Definition 8 (Restricted isometry property)**.** A linear operator $\Omega : \mathbb{R}^{n\times n} \to \mathbb{R}^s$ satisfies p-RIP with constant δ if for all $X \in \mathbb{R}^{n\times n}$ of rank at most p,

$$(1 - \delta)\|X\|_F^2 \leq \|\Omega(X)\|^2 \leq (1 + \delta)\|X\|_F^2 .$$

This definition encodes the notion that Ω preserves the norm of low-rank matrices under its transformation. We can prove a simple lemma that extends this to the inner product.

Lemma 24. *If Ω is $(p+q)$-RIP with parameter δ, then for any symmetric matrices X and Y of rank at most p and q respectively,*

$$\Omega(X)^T\Omega(Y) \geq \mathbf{tr}\,(XY) - \delta\|X\|_F\|Y\|_F$$

Proof. For any $a \in \mathbb{R}$, since Ω is linear,

$$\begin{aligned}\mathbf{tr}\left(\Omega(X)\Omega(Y)\right) &= \frac{1}{4a}\left(\|\Omega(X)+a\Omega(Y)\|^2 - \|\Omega(X)-a\Omega(Y)\|^2\right)\\ &= \frac{1}{4a}\left(\|\Omega(X+aY)\|^2 - \|\Omega(X-aY)\|^2\right).\end{aligned}$$

Since $\mathbf{rank}\,(X-aY) \leq \mathbf{rank}\,(X) + \mathbf{rank}\,(Y) \leq p+q$, we can apply our RIP inequalities, which produces

$$\begin{aligned}\mathbf{tr}\left(\Omega(X)\Omega(Y)\right) &\geq \frac{1}{4a}\left((1-\delta)\|X+aY\|_F^2 - (1+\delta)\|X-aY\|_F^2\right)\\ &\geq \frac{1}{4a}\left(-2\delta\|X\|_F^2 + 4a\mathbf{tr}\,(XY) - 2\delta a^2\|Y\|_F^2\right)\\ &= \mathbf{tr}\,(XY) - \delta\frac{\|X\|_F^2 + a^2\|Y\|_F^2}{2a}.\end{aligned}$$

Substituting $a = \frac{\|X\|_F}{\|Y\|_F}$ results in

$$\mathbf{tr}\left(\Omega(X)\Omega(Y)\right) \geq \mathbf{tr}\,(XY) - \delta\|X\|_F\|Y\|_F,$$

as desired. □

Finally, we prove our main theorem that shows that the quadratically transformed objective function is strongly convex in a ball about the solution.

Theorem 4. *If we define $f(Y)$ as the objective function of the above optimization problem, that is for $Y \in \mathbb{R}^{n\times p}$ and $A \in \mathbb{R}^{n\times n}$ symmetric of rank no greater than p,*

$$f(Y) = \left\|\Omega(YY^T - A)\right\|^2,$$

and Ω is $3p$-RIP with parameter δ, then for all Y, if we let λ_p denote the smallest positive eigenvalue of A then

$$\nabla_V^2 f(Y) \succeq 2\left((1-\delta)\lambda_p - (3+\delta)\left\|YY^T - A\right\|_F\right)I.$$

Proof. The directional derivative of f along some direction V will be, by the product rule,

$$\nabla_V f(Y) = 2\Omega(YY^T - A)^T\Omega(YV^T + VY^T).$$

The second derivative along this same direction will be

$$\begin{aligned}\nabla_V^2 f(Y) &= 4\Omega(YY^T - A)^T\Omega(VV^T) + 2\Omega(YV^T + VY^T)^T\Omega(YV^T + VY^T)\\ &= 4\Omega(YY^T - A)^T\Omega(VV^T) + 2\left\|\Omega(YV^T + VY^T)\right\|^2.\end{aligned}$$

To this, we can apply the definition of RIP, and the corollary lemma, which results in

$$\nabla_V^2 f(Y) \geq 4\mathbf{tr}\left((YY^T - A)(UU^T)\right) - 4\delta\left\|YY^T - A\right\|_F\left\|UU^T\right\|_F + 2(1-\delta)\left\|YU^T + UY^T\right\|_F^2.$$

By Cauchy-Schwarz,

$$\begin{aligned}\nabla_V^2 f(Y) &\geq -4\left\|YY^T - A\right\|_F\mathbf{tr}\left(UU^T\right) - 4\delta\left\|YY^T - A\right\|_F\mathbf{tr}\left(UU^T\right) + 2(1-\delta)\lambda_{\min}(Y^TY)\mathbf{tr}\left(UU^T\right)\\ &= 2\left((1-\delta)\lambda_{\min}(Y^TY) - 2(1+\delta)\left\|YY^T - A\right\|_F\right)\mathbf{tr}\left(UU^T\right).\end{aligned}$$

Now, since at the optimum, $\lambda_{\min}(Y^T Y) = \lambda_p$, it follows that for general Y,

$$\lambda_{\min}(Y^T Y) \geq \lambda_p - \left\| YY^T - A \right\|_F .$$

Substituting this in to the previous expression,

$$\begin{aligned} \nabla_V^2 f(Y) &\geq 2\left((1-\delta)(\lambda_p - \left\| YY^T - A \right\|_F) - 2(1+\delta)\left\| YY^T - A \right\|_F\right) \mathbf{tr}\left(UU^T\right) \\ &= 2\left((1-\delta)\lambda_p - (3+\delta)\left\| YY^T - A \right\|_F\right) \|U\|_F^2. \end{aligned}$$

Since this is true for an arbitrary direction vector U, it follows that

$$\nabla_V^2 f(Y) \succeq 2\left((1-\delta)\lambda_p - (3+\delta)\left\| YY^T - A \right\|_F\right) I,$$

which is the desired result. □

This theorem shows that there is a region of size $O(1)$ (i.e. not dependent on n) within which the above problem is strongly convex. So, if we start within this region, any standard convex descent method will converge at a linear rate. In particular, coordinate descent will do so. Therefore, we can imagine doing the following:

- First, use Alecton to, with high probability, recover an estimate Y that for which $\left\| YY^T - A \right\|_F$ is sufficiently small for the objective function to be strongly convex with some probability. This will only require $O(n \log n)$ steps of the angular phase of the algorithm per iteration of Alecton, as stated in the main body of the paper. We will need p iterations of the algorithm to recover a rank-p estimate, so a total $O(np \log n)$ iterations will be required.

- Use a descent method, such as coordinate descent, to recover additional precision of the estimate. This method is necessarily more heavyweight than an SGD scheme (see Section E for the reason why an SGD scheme cannot achieve a linear rate), but it will converge monotonically at a linear rate to the exact solution matrix A.

This *hybrid method* is in some sense a best-of-both worlds approach. We use fast SGD steps when we can afford to, and then switch to slower coordinate descent steps when we need additional precision.

Secondary Literature

[11] Emmanuel Candès, Xiaodong Li, and Mahdi Soltanolkotabi. Phase retrieval via wirtinger flow: Theory and algorithms. *arXiv preprint arXiv:1407.1065*, 2014.

[13] EmmanuelJ. Candès and Xiaodong Li. Solving quadratic equations via phaselift when there are about as many equations as unknowns. *FoCM*, 14(5):1017–1026, 2014.

[28] R.H. Keshavan, A. Montanari, and Sewoong Oh. Matrix completion from a few entries. *Information Theory, IEEE Transactions on*, 56(6):2980–2998, June 2010.

[41] Praneeth Netrapalli, Prateek Jain, and Sujay Sanghavi. Phase retrieval using alternating minimization. In *Advances in Neural Information Processing Systems 26*, pages 2796–2804. 2013.